Investigating the construct of productive vocabulary knowledge with Lex30.

Clenton, Jonathan

Investigating the construct of productive vocabulary knowledge with Lex30.

Jonathan Clenton

Submitted to Swansea University
in fulfilment of the requirements
for the Degree of
Doctor of Philosophy

Swansea University

2010

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Summary

This thesis investigates the construct of productive vocabulary knowledge with a productive vocabulary task, Lex30. The task is designed to elicit up to four vocabulary items in response to each of 30 cues. In this way, Lex30 generates a corpus for each subject up to 120 words, which is then categorised according to frequency bands. The test is scored according to the number, or proportion, of infrequent words elicited, with infrequent defined as all items excluding the most frequently occurring 1000 English words. The higher the Lex30 score, then, the more infrequent words that subject has produced in response to the cues. Each corpus generated by Lex30, therefore, offers information about subjects' relative knowledge of infrequent items, although this might only be threshold knowledge.

A feature of Lex30 is that it appears to measure productive vocabulary knowledge discretely: it does not activate multiple aspects of language knowledge, and is not context engaging. This feature suggests that we can measure one of the many aspects that are commonly considered to constitute language knowledge, productive vocabulary knowledge, without interference from other aspects of language knowledge. Additionally, Lex30 offers the potential to hypothesize about subjects' relative L2 proficiency in terms of the proportion of infrequent items they provide.

To investigate the construct of productive vocabulary with Lex30, this thesis examines, in a principled way, exactly what aspect of language competence it measures, and makes comparisons with other cognate tests. The test has been used in a number of contexts since its introduction; this thesis offers a thorough investigation of its reliability, different versions of the scoring system, the influence cue frequency and of specific cue items, and the mode of task delivery and response. The thesis concludes that Lex30 provides us with a helpful means to understand the construct of productive vocabulary knowledge.

DECLARATION This *work has not previously been accepted for any degree and is not currently being submitted in candidature for any degree.*

STATEMENT 1

This thesis is the result of my own investigation except where otherwise stated. Other sources are acknowledged by explicit references. A bibliography is appended.

SIGNED.....

DATE 1st August 2010

STATEMENT 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan and for the title and summary to be made available to outside organisations.

SIGNED.....

DATE 1st August 2010

Table of Contents

Acknowledgements. I would like to thank my supervisor, Dr. Tess Fitzpatrick, for her invaluable direction and guidance throughout the completion of this thesis. Very many thanks also go to Professor Paul Meara. I would also like to thank my colleagues at Osaka University for their kind help.

List of Tables and Figures

Chapter 1 Introduction

1.1 Vocabulary Knowledge.

This thesis investigates the construct of productive vocabulary knowledge with a test of productive vocabulary, Lex30. Before beginning to evaluate Lex30 and other measures of productive vocabulary knowledge I shall briefly explore what vocabulary knowledge implies. In short, vocabulary knowledge is difficult to define succinctly, and once we begin to consider what subjects' vocabulary knowledge might consist of, an understanding of such knowledge becomes multifaceted. The many aspects of knowledge that are potentially included in 'knowing a word' complicate this issue. Nation's (2001: 27) table of what is involved (shown in table 2.1, p.24, and discussed in detail in chapter two and chapter eight) highlights the complexity of this issue. For instance, Nation's aspects of knowledge fit into three groups: knowing the form of a word (which consists of both receptive and productive knowledge of its spelling, sound, and word parts); knowing the meaning of a word (which consists of both receptive and productive knowledge of its form and meaning, knowing a concept for the word and what it can refer to, and knowing what other words of related meaning it can be associated with); and knowing how a word is used (which consists of both receptive and productive knowledge of the grammar of the word including the part of speech and the sentence patterns it fits into, collocates of the word, and whether the word is formal or informal). In addition, Laufer (2005) also lists aspects learners need to bear in mind when stating that they have mastery of a vocabulary item, including: form, word structure, grammatical features, verb patterns, different meaning types, and so on. Thus, both Nation's and Laufer's facets of vocabulary knowledge highlight the potential impracticality involved in measuring these various aspects of vocabulary knowledge.

The distinction Nation, and many others (Laufer 1998, Read, 2000, Schmitt and McCarthy 1997, Webb 2005, 2007) make between receptive and productive knowledge should be mentioned to carry this last point further. Receptive knowledge is sometimes

called passive knowledge and is what is needed to deal with vocabulary when listening and reading. Productive vocabulary, sometimes called active knowledge, is needed when speaking and writing. The amount of contextual information required to listen or read suggests that receptive, or passive, vocabulary knowledge is greater than productive or active vocabulary knowledge. The difficulty with receptive and productive vocabulary knowledge is that there is a distinct lack of clarity as to whether they are separate or related aspects of knowledge (Melka 1997: 84). Melka (1982: 6-7) proposes that the terms ought to be considered in terms of degrees of knowledge, in terms of a continuum, in which words that are understood receptively become available for productive use (1983: 21). Melka (1982) suggests that the 'boundaries' of receptive and productive vocabulary vary and are dependent on each subject and environmental demands. In any case, discussion related to the distinction between receptive and productive vocabulary knowledge serves to highlight the complexity and scope of any attempt to describe vocabulary knowledge. Yet, it is first important to realize, as Read (1988: 35) suggests, what areas of lexical competence we as researchers and language teachers want to test.

1.2 Testing productive vocabulary knowledge.

There are numerous interrelated aspects of vocabulary knowledge as our brief summary of Nation's table (2001: 27) in the previous section and Laufer's (2005) facets of word knowledge show. Knowledge of a word is very difficult to measure since knowledge of one aspect of vocabulary implies knowledge of other, related, aspects. For example, if a subject is able to produce a particular word when prompted, should we take this to mean that they know other related aspects? When a test sets out to elicit a particular word from a subject, and if that subject produces the particular word, we ought to ask how we can assume that their knowledge of that word reflects knowledge, amongst others, of the word morphology, its use in different contexts, and potential meanings. In addition, in the eliciting of that word, we might ask whether we have assumed knowledge of other words. If a test score tells us that 'Tanaka san' *knows* approximately 2400 word families

in English, the extent to which Tanaka san knows each of these word families obviously varies from word to word. For instance, Tanaka san might be very familiar with the written form of most of the words at her disposal but may not be comfortable, or familiar, with the spoken forms. Alternatively, Tanaka san might be more familiar with some of the words acquired at the earlier stages of learning, and be aware of many contexts in which to apply those words, yet when asked to apply knowledge of more recently learned items might face greater difficulty. Thus, the issue here is whether tests fully reflect the nature of a subject's vocabulary knowledge. A further and potentially more fundamental concern is that subjects obviously vary in terms of their ability to apply their individual knowledge, and the extent to which they respond consistently to particular test types (even though they may know the same number of words as other better performing students) (Laufer 2005: 584), or in terms of their depth of knowledge of very similar set of vocabulary items. The difficulties raised here represent the multi-dimensional and multi-faceted nature of vocabulary knowledge.

As well as considering the many aspects of related knowledge word knowledge implies, we also need to disentangle what we aim to measure in terms of the unit we seek to assess. We could count all of the different types subjects know (such as the different forms of the verb break: *break, breaks, broke, broken*). The lemma of *break* though is comprised of the root, *break,* with the most frequent regular inflections. Hence *describe* and *describes* form a single lemma *describe.* Alternatively, we could describe vocabulary knowledge in terms of *word families,* which are more inclusive than lemma in the sense that they include all possible permutations of the vocabulary type. Thus, the word family for break would include *break, breaks, broke, broken, breakage, unbreakable, breaking* and so on, whereas the lemma would only be *break.*

These difficulties may be complicated further once we attempt to assess productive vocabulary. Nation's aspects are a useful starting point for discussing what needs to be accessed when attempting to test knowledge of productive vocabulary. Although before

dealing with assessment, it is crucial to consider what we mean by productive vocabulary. Vocabulary is produced in either a written or a spoken form and while I write this thesis, the words I produce are those that relate to the other words in the sentences, which suggests that there is a semantic component to my productive vocabulary knowledge. Regardless of the 'depth' (or the degree) of knowledge I might have of the items I produce, they are still reflective of my productive vocabulary knowledge. Thus, I can say that I am able to produce words while I write this thesis and can argue that these are being tapped from my productive vocabulary knowledge.

In order to tap productive vocabulary knowledge other aspects of knowledge have to be accessed. For instance, I know that when I ask an L2 learner to tell me all of the L2 words they know beginning with the letter 'A', for example, in response to Spreen and Benton's (1977) Controlled Oral Word Association Test, they need to understand my L2 directions. If they cannot follow the directions, they cannot produce the words. Yet the direction clearly necessitates that other aspects of knowledge are accessed in this case when I attempt to tap my subject's productive vocabulary knowledge. To return to Nation's table (2001: 27), the direction (when I ask them to 'provide all the words they know beginning with the letter 'A'') includes receptive knowledge of, for instance, form (what do the directions/ words sound like?), word parts (what parts are recognizable in these directions/ words?), form and meaning (what meaning do these word forms/ directions signal?) and so on. I could write the directions, but this would still imply that receptive knowledge is tapped when my subjects attempt to decipher the demands for the task. From this brief example, it is obvious that other aspects are involved by even a simple task designed to elicit productive vocabulary.

By comparing different responses to such a word association task we might then speculate that the kinds of responses provide, tell us something about the learners' productive vocabulary knowledge. This might take the form of speculating about the frequency or infrequency of the words provided (Nation 1984), the extent to which the

words might be accessed or activated by knowledge of other words (Meara 2006: 625), or lead us to query whether the learner can either 'use' or only 'recall' (Read 2000:156) the items.

In terms of vocabulary assessment, it is therefore apparent that there are numerous interrelated factors to be considered in testing. For instance, the unit of measurement needs to be decided upon, as well as the need to ensure that the assessment is appropriate for the intended test. In this second sense, any test of productive vocabulary is likely to be a multi-faceted entity because of the fact that it cannot *only* include aspects of productive vocabulary knowledge, because in order to access productive vocabulary knowledge, subjects' receptive vocabulary knowledge must also be accessed as the above section indicates. Subjects should also be assessed fairly and evenly in spite of the likely variations in terms of their own preferences for certain types of test. Having retrieved some items, we might then begin to speculate that learner responses convey something about each individual's productive vocabulary knowledge.

1.3 Tests may not be measuring what they claim.

As the above section shows, a potential problem with testing is that numerous aspects of vocabulary knowledge are being accessed at the same time. If the aim is to examine and assess productive vocabulary knowledge, then the assessment should access just that with minimal recourse to other aspects of vocabulary knowledge. If other aspects of vocabulary knowledge are accessed then this might muddy claims that the assessment only measures productive vocabulary knowledge. If we are to measure a particular aspect of vocabulary, we then need to do so with minimal recourse to other aspects of knowledge. This might also justify our claims to have largely accessed the trait we are trying to measure, which is important in order to claim that the intended measurement has accessed what we claim it has, but also to separate and potentially identify the individual strands that make up the construct of lexical competence. We have already

seen two attempts (Nation 2001, Laufer 2005) to describe vocabulary knowledge. Yet we have little in the way of confirmation that vocabulary knowledge consists of only or all of these separate aspects of knowledge. Thus, tests might not measure what they claim to. Once we concede that tests might be accessing multiple aspects of knowledge in their assessments, we then ought to admit that they may not tell us very much about the traits they claim to measure. If tests hide subjects' abilities then they might not be able to say very much about the knowledge they should be measuring. If tests measure discrete and separable aspects of knowledge, without accessing multiple aspects, then we can be far more confident about the definitions of the aspects such tests claim to measure.

1.4 Lex30.

Meara and Fitzpatrick (2000: 28) claim to have designed a test that elicits productive vocabulary with some of the factors discussed above in mind. Lex30, they claim, is a test that does not make any assumptions about the relationship between receptive and productive vocabulary knowledge. Thus, the issues discussed above raise two important concerns that Lex30 might begin to resolve. First, we need to ensure that a measure of productive vocabulary knowledge accesses only minimal other aspects of vocabulary knowledge and minimises the effect of these on the score. Second, if Lex30 does access productive vocabulary with only minimal recourse to other aspects of vocabulary knowledge, we should see how accurately and consistently the test works. In short, Lex30 might access productive vocabulary with minimal recourse to other aspects of knowledge; so the extent to which Lex30 accesses representative samples of productive vocabulary needs examining.

The next chapter (2) of the thesis presents the literature review, which is divided into three sections. The first section evaluates Meara and Fitzpatrick's claim that Lex30 "might tap the extent of non-native speakers' productive vocabulary more effectively than some other tests in current use" (2000:19) and does this by surveying the studies

that claim to evaluate productive vocabulary. The second section does this by evaluating Lex30. The third and final section examines attempts to describe the construct of lexical competence. The experimental chapters (3-7) assess Lex30 as a test of productive vocabulary knowledge. Accordingly, the thesis will attempt to explore the extent to which Lex30 is able to tap into the construct of productive vocabulary knowledge, and to determine the role Lex30 has in understanding productive vocabulary in relation to other lexical studies.

Chapter 2 Literature Review

2.1 Introduction.

In order to investigate the construct of productive vocabulary and the effectiveness of Lex30, the extent to which other existing measures access productive vocabulary needs evaluating. My aim is to examine whether the construct of productive vocabulary is accessed and measured by the tests reviewed in this chapter (2), and to provide a platform for the experimental chapters (3-7) that follow on from this literature review.

The literature review is divided into three sections. The first section presents a review of nine studies all chosen as attempts to assess productive vocabulary knowledge. While none of the studies presented in these papers exclusively set out to access productive vocabulary knowledge all relate to the general field of study. I should also add that these tasks were not designed specifically to investigate productive vocabulary knowledge of learners, but, for example, to test vocabulary learned as part of a course (Wesche and Paribakht's 1996 Vocabulary Knowledge Scale) or to determine productive vocabulary produced in a composition task (Laufer and Nation's 1995 Lexical Frequency Profile).

The second section in this review discusses papers related to the Lex30 task, which, as mentioned in chapter one, though primarily designed as a research tool, aims to elicit and measure productive vocabulary.

The third and final section collects strands of the discussions from sections one and two by discussing two papers that summarize attempts to define the construct of lexical competence to see how these reviews help to identify the construct of productive vocabulary. The end of this literature review presents a summary of the main points of all three sections and closes by suggesting the need for further detailed experimental examination of the claims made throughout this chapter.

2.1.1 Productive vocabulary – not a straightforward construct.

Before presenting a more detailed summary of the papers in this review, it should be emphasized that productive vocabulary is not a simple or straightforward construct. In Nation's (2001) summary of 'what is involved in knowing a word' (table 2.1 below) each aspect of knowledge includes both a receptive and a productive facet. From the outset, this shows how difficult it may be to say, with any degree of confidence, that any of the tests in this review measures productive vocabulary in isolation.

Table 2.1 What is involved in knowing a word? (Nation, 2001:27).

Form	Spoken	R	What does the word sound like?
		P	How is the word pronounced?
	Written	R	What does the word look like?
		P	How is the word written and spelled?
	Word parts	R	What parts are recognizable in this word?
		P	What word parts are needed to express the meaning?
Meaning	Form and meaning	R	What meaning does this word form signal?
		P	What word form can be used to express this meaning?
	Concept and referents	R	What is included in the concept?
		P	What items can the concept refer to?
	Associations	R	What other words does this make us think of?
		P	What other words could we use instead of this one?
Use	Grammatical functions	R	In what patterns does the word occur?
		P	In what pattern must we use this word?
	Collocations	R	What words or types of words occur with this one?
		P	What words or types of words must we use with this one?
	Constraints on use	R	Where, when, and how often would we expect to meet this word?
	(register, frequency...)	P	Where, when, and how often can we use this word?

Note: In column 3: R = Receptive vocabulary knowledge; P = Productive vocabulary knowledge

2.2 Measures related to the construct of productive vocabulary knowledge.

A fundamental issue that first needs tackling is whether productive vocabulary knowledge can be separated from other aspects of language knowledge. A review of the measures in section 2.2 suggests that the more the assessment of productive vocabulary knowledge is contextualized, the greater the possibility of other aspects of language knowledge coming into play. By presenting subjects with a sentence or composition completion task, for instance, subjects might have to call upon productive as well as receptive knowledge (such as knowledge of grammar or collocations), or indeed other aspects, such as those Nation (2001: 21) refers to in his summary (table 2.1). The issue here is that the greater the context provided by certain tests the less we can be certain that only productive knowledge is being evaluated.

The range of studies covered by this review indicates that the subject area is far from straightforward and requires explication. The nine papers discussed in this first section are reviewed in order to determine what they measure and to clarify what the authors mean by productive vocabulary knowledge. The section starts with a review of Wesche and Paribakht's (1996) 'Vocabulary Knowledge Scale' (VKS), which addresses both receptive and productive knowledge. Laufer and Nation's (1999) much cited 'controlled test for productive vocabulary' follows and highlights the difficulty of eliciting knowledge from sentences. A review of Laufer and Paribakht (1998) then follows, in which they attempt to distinguish between two different aspects of productive vocabulary namely 'controlled productive' and 'free productive vocabulary'. Laufer *et al.* (2004) examine the issue of size and use with their Computer Adaptive Test of Size and Strength (CATSS) used in assessing knowledge of vocabulary meaning. This theme is picked up on in discussions of the two papers by Webb. Webb attempts to test vocabulary knowledge by using nonsense words, thereby attempting to guarantee that there is no prior knowledge of the items being tested. Webb's (2005 and 2007) measures are revealing since they claim to isolate the influence of context in testing productive knowledge and separate out different aspects of knowledge. The final measure reviewed

in this section is Laufer and Nation's (1995) Lexical Frequency Profile (LFP). The LFP sees a departure from what Read describes as (2000: 115) 'discrete and selective' tests with this 'embedded, comprehensive and context dependent' measure. Meara's (2005) paper presents a critique of Laufer and Nation's LFP which Laufer (2005) responds to. These papers from Meara (2005) and Laufer (2005) have implications for the creation of a test to access the construct of productive vocabulary, and they force an examination of whether productive 'use' is an accurate indicator of the size of a subject's productive vocabulary.

2.2.1 Wesche and Paribakht (1996): Assessing second language vocabulary knowledge: Depth vs. Breadth.

A review of Wesche and Paribakht's Vocabulary Knowledge Scale (VKS) highlights the much-cited distinction between breadth and depth of vocabulary knowledge (Anderson and Freebody 1981, Henriksen 1999, Laufer and Paribakht 1998, Laufer 1998, Read 2000, Read 2004). The VKS attempts to measure the 'quality,' or depth, of knowledge (of a specific number of words (24)) while much of the work of vocabulary researchers has been dominated by the search for a learner's vocabulary 'size', or breadth (such as Laufer 1992, 1998; Laufer and Nation 1995, 1999; Laufer *et al.* 2004, Meara 1996a; Meara and Fitzpatrick 2000; Nation 1983, 1990; Read 1988, 2000).

Wesche and Paribakht assessed a subject group of 17 learners twice over two weeks with the VKS. Twenty-four target words were selected from the subjects' own course material. The 24 words were selected from course themes from an undergraduate program for non-native students at the University of Ottawa. The test took the form of subjects self-reporting their knowledge on the following 5-point scale:

1. I don't remember having seen this word before
2. I have seen this word before, but I don't know what it means

3. I have seen this word before and I think it means _____
4. I know this word. It means _____
5. I can use this word in a sentence. E.g.:_____

Wesche and Paribakht (1996: 32).

Wesche and Paribakht describe how an accurate answer to level 5 represents productive knowledge while levels 2, 3, and 4 are measures of recognition and level 1 indicates no knowledge of a word. Subjects gain credit for evidence of any knowledge provided in response to questions 3, 4 and 5. A corresponding mark is awarded for each appropriately answered question, hence 5 points is awarded for question 5, 4 points for 4 and so on. Partial credit is awarded, for instance, when a correct response is given to a question but is not backed up by appropriate evidence in response to 5, and so on.

Wesche and Paribakht found significant and strong correlations between the scoring and the subjects' self-assessments (0.92 to 0.97) and suggest that the method of elicitation was reasonably reliable. Wesche and Paribakht also found that the test scores indicate that subjects operate in broadly the same way over the two different test periods examined. They found low correlations in their concurrent validity test (0.53, $p<.01$) with the Eurocentres Vocabulary Size test (EVST) (Meara and Jones 1990) (described in detail in section 2.3.1). These low correlations may be unsurprising since the 24 words tested may not have been representative of the subjects' lexicons.

Wesche and Paribakht's paper, which has been widely cited (e.g. Henriksen 1999, Laufer and Paribakht 1998, Meara and Wolter 2004, Read 2004, Wesche and Paribakht 1999, Wolter 2001), is noteworthy since it was the first test of vocabulary to incorporate the idea of incremental knowledge. However, in terms of measuring productive vocabulary Wesche and Paribakht's (1996) VKS is problematic. By eliciting responses in the form of sentences in the way that they do, Wesche and Paribakht appear to be placing too great a burden on their examiners. This potential issue relates to responses to

level 5, in which subjects are rewarded for providing semantically appropriate and accurate responses. It is far from clear how examiners might award credit for when subjects provide different sentences as examples of their knowledge. If we consider a word such as 'established,' by way of an example, there is likely to be subject variation within the subjects' responses (e.g., providing such responses, with the word 'established' as 'It has been established', 'the answer to his question has been established' or 'I thought his answer to the question has been established since it is known by...' etc). With no explicit indication of what would be considered an acceptable response to level 5, and 3 or 4, there is no way of knowing how these different responses might be dealt with by examiners. Thus, the call upon examiners to interpret subjects' responses is demanding and open to misinterpretation.

Accordingly, while subjects may be able to produce a semantically appropriate sentence such as 'I understand that the question is relevant' as evidence of the ability to use the word 'relevant' this may leave us with no real understanding of their productive knowledge of the item elsewhere other than in the lexical phrase subjects provide. Even though Wesche and Paribakht (p.33) propose an additional scale with a sentence along the lines of 'the following includes all the meanings I know for the word_____' this still provides subjects with an unnecessarily laborious task. Producing all the meanings one knows of a word is especially difficult, because words are not usually remembered until they are needed. In either case, by requiring subjects to respond in this way, the VKS elicits responses that have the potential to be interpreted ambiguously and, for this reason, examiners may or may not be crediting knowledge accurately and consistently.

This criticism aside, it is worth remembering that Wesche and Paribakht's VKS was originally designed to test vocabulary their student subjects were expected to learn as part of their course. While Wesche and Paribakht did not set out to test productive vocabulary exclusively, their test provides us with several useful considerations relating to how to access productive vocabulary. Thus, sentence completion tasks might not accurately reflect a subject's knowledge of an item needs to be considered in view of the

likely reliance on examiners' interpretations in scoring subject's responses. This appears to be problematic, because subjects' written responses are clearly open to misinterpretation.

Wesche and Paribakht's VKS task suggests that productive vocabulary knowledge is demonstrated in relation to the ability to use the items providing an L2 example sentence. They only address productive vocabulary in level five of their task at which they imply that productive knowledge of an item is demonstrated by providing an example sentence. At this level, subjects might activate other, possibly multiple, aspects of linguistic knowledge aside from productive vocabulary knowledge. Yet Chapelle (1998: 43) highlights the difficulty and inseparability of such a perspective with examples of the kinds of L2 knowledge that might be activated, such as 'sentence context' , 'context of language use', 'level of formality', as well as the need to know 'which words' are necessary'.

2.2.2 Laufer and Nation (1999): A vocabulary-size test of controlled productive ability.

Laufer and Nation's controlled productive knowledge test addresses some of the problems with Wesche and Paribakht's VKS by narrowing the scope of their sentence task. Laufer and Nation's 'controlled productive ability' test assesses the ability to produce a word when primed to do so by the first few letters of a target item in a context sentence. Laufer and Nation's test was born out of the success of Nation's Vocabulary Levels Test (1983) and aimed to be an equivalent test of productive knowledge. Their test sampled 18 items at each of the 2k, 3k, 5k, UWL (University word list), and the 10k word levels in the same way as the original levels test.

In their controlled productive ability test, Laufer and Nation (1999) provided the first letters and sentence context of target words and required subjects to complete the words at each of the five levels tested. Laufer and Nation suggested that in this way testees are prevented from responding with other, albeit semantically appropriate items, from

different frequency levels. The authors tested their subjects with three parallel test versions using alternative vocabulary items from three parallel versions of the Vocabulary Levels Test. Thus, subjects were tested with items selected from the same frequency bands but with different items for each of the three parallel test versions.

Laufer and Nation carried out two studies, one to test the reliability and validity of one version, and a second to test the equivalence of three parallel forms of the Vocabulary Levels Test. In their first study, for validity, Laufer and Nation wanted to establish whether their test of 'controlled productive vocabulary' distinguished between different levels of language proficiency. Accordingly, their study was designed to elicit whether subjects performed differently from each other in terms of their overall test scores, and in terms of their scores within each of the different vocabulary levels (at each of the five frequency bands: 2k, 3k, the University Word List (UWL), 5k, and 10k). Thus, the controlled productive ability test was given to four different groups. The groups were differentiated in terms of their respective English (L2) proficiency levels. The level for each group was decided based on the length of L2 study ranging from the tenth grade at high school to first year university students. Subjects scored a point for each correct response they provided, such as 'connects' as in 'The railway con_____ the city with its suburbs'. The subjects were given six scores as follows: a score for the number of correct items at each of the (2k, 3k, UWL, 5k, and 10k) word levels and a score for the total number of correctly retrieved items. Laufer and Nation's results suggest that the subjects' scores improve with proficiency. They also found that the four groups' performances at each of the subsequent frequency band levels decreased. They extrapolate that the results demonstrate that subjects have mastered a word at each successive frequency level if they are able to complete the test sentences correctly (as in the example 'connects' above). Laufer and Nation interpreted this as evidence of test validity for their measure of vocabulary.

In their second study, Laufer and Nation devised three alternative versions of their original test using different items from the same frequency bands. They wanted to

determine whether subjects' scores on the four versions of the test would correlate highly with one another. Each subject took four different versions of a particular proficiency level. In other words, one group of learners sat four versions of the 2000-word level, while another group sat four versions of the 3000-word level, and so on. Table 2.2 below shows that the correlations were moderate to high between the different versions of the frequency band test. The correlations between the 5000 level tests were likely to be weaker than the other levels because the words came from a greater pool (i.e. 18 items were selected from 4999 infrequent words (5001 to 10,000) compared to, for example, 18 items from more frequent 2000 words for the 2k level). Laufer and Nation conclude "the Productive Vocabulary Levels Test is a reliable, valid and practical measure of vocabulary growth" (1999: 44).

Table 2.2 Correlations between four versions of the Productive Vocabulary Levels Test at four of the five frequency levels Laufer and Nation (1999: 43).

Level (n)	A/B	A/C	A/D	B/C	B/D	C/D
2000 (n45)	.82*	.82*	.78*	.83*	.81*	.77*
3000 (n36)	.71*	.70*	.82*	.82*	.71*	.80*
UWL (n 33)	.75*	.80*	.84*	.83*	.76*	.80*
5000 (n 18)	0.72 (p = .004)	.83*	0.69 (p = .003)	0.49 (p = .1)	0.77 (p = .003)	0.67 (p = .006)

Note: *significant at .0001 level

As with Laufer and Paribakht's (1998) paper, reviewed next (2.2.3), the authors' findings suggest that knowledge of 15 of the 18 items tested is sufficient to indicate mastery of a particular frequency band. Laufer and Nation claim to have designed a test that provides evidence of particular vocabulary knowledge amongst learners in terms of the five frequency levels assessed. Laufer and Nation argued that their test constrains subjects to provide a particular word for a given context and that the ability to complete this task successfully indicated knowledge of a particular frequency band.

Although this test has been widely used (Nation 2001, Read 2000, Read and Chapelle 2001, Vermeer 2001) and produces broadly meaningful results, there are a number of problems inherent in its design. One is that subjects are only tested with a limited set of items for a whole frequency band. As there is no guarantee that the contexts provided adequately reflect the frequency band being tested, there is no way of knowing how fair the test is. Although the sentence "Her beauty and cha____ had a powerful effect on men" (Laufer and Nation 1999: 46) might call upon collocational knowledge (of 'beauty and charm'), one would expect that this kind of collocational knowledge might evolve in subjects with a greater lexical base than 2000 items. Subjects might equally respond with "chatter", "chaise longue", or "charisma" for which they would not get credit, but which would reflect knowledge of items beyond the 2k level. There is also no way of knowing whether the sentence tasks are reflective of the subjects' own vocabulary development. Thus, testing with a sentence such as "the telegram was del_____ two hours after it had been sent" may present problems for subjects without an understanding of 'telegram' compared to the more modern application of, say, 'email.' In either of these instances, the test may require knowledge beyond or separate from the frequency level of the ('controlled') productive vocabulary supposedly being elicited.

Laufer and Nation's measure is not testing productive vocabulary knowledge exclusively. Subjects are required to call upon receptive knowledge in order to confirm whether 'productive' vocabulary items might fill a gap appropriately. Subjects need to have an understanding of the parts of speech of the surrounding items within each of the test sentences, as well as an understanding of the part of speech of the item the test sentence aims to elicit. For instance, if I am to understand that the target word 'delivered' is required in "the telegram was del_____ two hours after it had been sent," I have done this based on my understanding of the items surrounding the target word. Additionally, subjects are being asked to access wider contextual knowledge in order to complete the test items. Thus, Laufer and Nation's (1999) 'controlled productive knowledge' test accesses more than productive vocabulary knowledge.

While the test claims to measure productive knowledge, there are obvious and multiple demands placed on the test takers. Subjects are required to demonstrate knowledge of the surrounding items presented in each test sentence. In order to have any idea of which particular target item is required, subjects need to have some degree of understanding of the surrounding context provided. Also, a greater burden is placed on subjects in cases where the surrounding items presented in each test sentence are less frequent than the target item they are expected to provide. Laufer and Nation's (1999) test confuses the attempt to access productive vocabulary knowledge since so much of the task requires aspects of knowledge (i.e. knowledge of collocations, semantic knowledge, etc.) that include receptive knowledge as well as productive knowledge. Without ensuring that subjects' productive knowledge is the focus of the test we cannot be certain that other aspects of knowledge influence each subject's performance on such a test.

Laufer and Nation's Productive Levels Test implies that knowledge of productive vocabulary can be assessed by a subject's ability to complete the words with the prompt provided. As with Wesche and Paribakht's VKS, the Productive Levels Test requires subjects to demonstrate knowledge of multiple aspects of knowledge in addition to productive vocabulary. This perspective of vocabulary knowledge stresses the inseparability of the assessment of productive vocabulary from other aspects of L2 knowledge. The perspective assumes that 'an individual [that] has developed strong L2 vocabulary knowledge (as revealed in..written assessments of this component), is ..[also] likely to have developed strong L2 grammatical skills or perhaps strong functional knowledge of the L2 in particular contexts" (Cummins, 2000: 123).

2.2.3 Laufer and Paribakht (1998): The relationship between passive and active vocabularies: Effects of language learning context.

Laufer and Paribakht's (1998) study introduces the 'Vocabulary Levels Test' (the Productive Levels Test) and the 'Lexical Frequency Profile' (LFP) and the study

compares results from these two tasks with the 'controlled productive knowledge' test reviewed above.

Laufer and Paribakht examined the relationships between three types of vocabulary knowledge with the same group of subjects. They used Nation's (1983, 1990) Vocabulary Levels Test to examine 'Passive' knowledge, the productive version of the Vocabulary Levels Test (The Productive Levels Test) (Laufer and Nation 1999, reviewed above (2.2.2)) to examine 'Controlled Active' knowledge, and the Lexical Frequency Profile (Laufer and Nation 1995) to examine 'Free Active' knowledge. Laufer and Nation's Free Active test (the Lexical Frequency Profile) introduced here is described below in section 2.2.6. This study, like Wesche and Paribakht's VKS (reviewed above in section 2.2.1) recognises that vocabulary knowledge is not a simple binary yes or no phenomenon.

The Vocabulary Levels Test was designed to test a subject's knowledge of words out of context. It consists of 90 items, made up of 18 'representative' words from five different frequency levels for which a potential maximum of 90 points are awarded. Each level tests knowledge of one of five frequency bands (2K, 3K, 5K, UWL, 10K). As in the following multiple-choice task, testees are asked to match target words with corresponding definitions such as:

1. copy
2. event _6_ end or highest point
3. motor
4. profit _3_ this moves a car
5. pity
6. tip _1_ thing made to be like another

Laufer and Paribakht (1998: 373).

The Controlled Active Vocabulary Test (reviewed in full above in 2.2.2) aimed to elicit target items from the five frequency bands in the form of the contextualized sentences.

For each target word, the first few letters were provided in order to eliminate other possibilities. The scoring maximum is the same as the Vocabulary Levels Test (90), as the five levels were tested with the same number of 18 lexical items. Subjects score if they produce the intended target item, correctly spelled. The Free Active (FA) Vocabulary Test (reviewed in section 2.2.6) requires its subjects to write a 300 – 400 word composition on one of two general topics. The composition is then taken as a sample of a learner's vocabulary knowledge in relation to the different frequency bands. Laufer and Paribakht report that while the relationship between controlled active and passive vocabulary types was inconsistent, learners with a larger passive vocabulary tended to have a larger controlled active vocabulary. The authors found that ESL learners only consistently displayed higher controlled active scores (than passive vocabulary scores), after 2 years residence in the foreign environment.

Laufer and Paribakht originally set out to determine whether any 'shifts' might occur in the three types of knowledge they examined and, if so, whether such shifts depended upon the learning contexts investigated. They examined their subjects at three different points over a seven to ten day period. Their paper allows us to compare different gains made using the three measures they apply. There are two significant flaws with the tests which suggest that their results should be treated with caution. The first relates to the recurring theme throughout the reviews in this section, that tests may not exclusively measure what the authors claim (though the authors (Wesche and Paribakht 1996; Laufer and Nation 1999; Laufer and Paribakht 1998) do acknowledge that the three tests might access different kinds of knowledge). The second relates to the assumed relationship between vocabulary size and test scores.

First, it is difficult to conceive of Laufer and Paribakht's 'controlled' productive knowledge test as testing only productive knowledge while multiple aspects of knowledge are being called upon. Subjects need to have at least a receptive or passive awareness of surrounding items in the sentence, as section 2.2.2 showed. Subjects also need an understanding of the part of speech of the item required (the simple present verb

'connects' for 'the railway con___ with its suburbs') to complete the sentence. The test is less a productive knowledge test of lexical items (where 'join' might otherwise be semantically acceptable) and more a semantic test requiring contextual knowledge. Laufer and Nation's task calls for specific vocabulary knowledge for gap fill tasks in which subjects might know other semantically equivalent items.

The second problem with Laufer and Paribakht's study is the assumption that there is a straightforward relationship between test scores and vocabulary size. They claim their 'results confirmed the general perception that learners' passive (P) vocabulary is larger than their controlled active (CA) vocabulary' (1998: 383). Yet this seems rather an ambitious assumption because the three tests clearly operate in different ways and make different demands on the subjects. The passive understanding test is a gap fill task offering a set of potential responses, while the controlled active test offers no such information and subjects need to provide the ending to a target word. In addition, only one of the tests is multiple choice, which might influence the way scores are obtained (i.e. allowing for guessing and an elimination process). Both tests require quite different aspects of passive and active knowledge and different test skills in order for subjects to achieve any degree of success. This theme of different test demands reappears in Laufer *et al.* (2004) reviewed in section 2.2.4 below.

For Laufer and Paribakht's (1998) Vocabulary Levels Test, the subjects are required to interpret the test sentences, which appears to place far too great a burden on test subjects when they are asked to provide evidence of their productive vocabulary knowledge. In addition, the controlled productive ability test might actually penalize subjects otherwise able to provide knowledge of semantically appropriate synonyms but are not presented with no opportunity to do so; subjects' vocabulary abilities are based only on the pre-selected items being tested. It is therefore too difficult to glean whether subjects' scores fully reflect the productive vocabulary knowledge they might otherwise have. Also, just how reliably eighteen sample targets might represent a frequency band remains unknown.

2.2.4 Laufer, Elder, Hill, and Congdon (2004). Size and strength: Do we need both to measure vocabulary knowledge?

Laufer *et al.* (2004), present a further claim to show a decontextualized test and in this case, one that measures different dimensions of word knowledge. The test Laufer *et al.* (2004) propose is one for meaning which, they suggest, overcomes the issue of having to account for every potential context in testing. The authors go on to propose that any such test should incorporate an understanding of meaning. They argue that a good vocabulary test should incorporate the extent to which test takers can correctly associate word form with the concept the form denotes. Laufer *et al.* point out that knowledge of meaning takes on many different forms and that, to date, not all researchers agree on a definition. In their attempts to avoid confusion, the authors distinguish four degrees of knowledge of meaning, based on the following two dichotomous distinctions:

- Supplying the form for a given concept vs. supplying the meaning for a given form; and
- Recall vs. recognition (of form or meaning).

The first distinction implies a difference between subjects who can retrieve the L2 word without a cue compared to those who can, once L2 options are presented and refer to these two abilities as 'active' and 'passive' knowledge. The second distinction implies a difference in knowledge between subjects who can recall the form or the meaning of a word and those who cannot, but can recognize the form or the meaning from a set of options. The following shows how each type of knowledge of meaning can be elicited for the same item ('melt', in this example):

Type 1: Active recall:	Turn into water m_____	
Type 2: Passive recall:	When something melts it turns into _____	
Type 3: Active recognition:	Turn into water:	(a) elect
		(b) blame

(c) melt

(d) threaten

Type 4: Passive recognition:	Melt:	(a) choose
		(b) accuse
		(c) make threats
		(d) turn into water

(Laufer *et al.* 2004: 206).

Laufer *et al.* call their test the Computer Adaptive Test of Size and Strength (CATSS). The adaptive element of the test arises if, for instance, a subject provides a correct response to the 'active recall' modality then the other modalities are not tested. If a subject guesses incorrectly, or fails to answer correctly, the word is presented in the next modality, and so on. As well as testing the size of a subject's vocabulary, Laufer *et al.* propose to test 'strength' of knowledge of meaning of a particular item. They distinguish strength from depth in the sense that depth incorporates supra meaning aspects such as pronunciation.

To test vocabulary 'size' the authors test subjects' knowledge of items that they claim to be representative of particular frequency levels. In total, thirty vocabulary items were randomly selected from five frequency levels (2k, 3k, 5k, 10k, and the Academic Word List) in the same way as with Nation's Vocabulary Levels Test (1983, 1990), Laufer and Paribakht (1998), and Laufer and Nation (1999). Vocabulary strength was tested according to the four modalities described above, moving from the most difficult, active recall, to the 'easiest,' passive recall.

Laufer *et al.* marked responses according to a key in which subjects received credit according to each modality and for each frequency level successfully completed. They suggested that the CATSS was an improvement on earlier testing in the sense that it

provided a more detailed account of subjects' scoring. They claim that their test is able to show that a subject might have a good vocabulary for reading at the 5000 frequency level, when scoring 27 out of 30 for passive recognition, but might have poor productive skills, perhaps scoring only 5 out of 30 for active recall.

Laufer *et al.* claim to have presented a decontextualized test that measures different dimensions of word knowledge. However, after closer examination, their methodology raises two significant problems requiring further attention. First, the success of their CATSS task depends on the word provided and assumes that subjects' lexicons are structured in similar ways. Second, they assume a cline from recall to recognition.

First, Laufer *et al.* claim that when presented with alternative modalities subjects may not be able to exhibit knowledge of a lexical item they were unable to provide knowledge of in an earlier 'modality' (p.204). However, in their CATSS test, if subjects are successful at guessing the most difficult modality first, active recall, then the remaining three modalities are not tested and the subject is assumed to know all the remaining modalities tested at that particular word frequency level. This is problematic since Laufer *et al.* assume subjects know the remaining three modalities in their hierarchy of aspects being tested (if subjects are able to demonstrate knowledge of active recall first). There is no way of confirming whether subjects are able to demonstrate knowledge of the alternative contexts of the item being tested or whether the surrounding L2 stimuli appropriately elicit the target item. The efficacy of the task must depend on the target item in question. In the example provided, 'melt,' it might be hard to imagine that subjects would not be able to answer the other three modalities correctly. Although there is no way of confirming whether this is likely to be the case for all lexical items being tested, it remains unknown how their four test sentences might elicit or confirm knowledge of more abstract terms such as 'elusive' or 'surreal'. The choice of four modalities might not comprehensively account for every potential meaning and might therefore fail to fully reflect subjects' productive vocabulary knowledge. Laufer *et al.*'s testing suggests that the hierarchy of a subject's lexicon is structured in the same

way for all items, along what is a cline from recall to recognition. Thus, Laufer *et al.*'s suggestion is somewhat problematic since there is no real way of validating whether subject's lexical knowledge is structured in this particular way.

Second, and despite Laufer *et al.*'s (2004) claims to the contrary, their test is not decontextualized. Subjects are required to demonstrate receptive understanding of the surrounding items in the test sentences provided. In addition, as in the criticism of Laufer and Paribakht (1998) above (2.2.3), Laufer *et al.* (2004) have tested their subjects with a pre-selected list of words that are then tested in pre-ordained sentences. Unless the learning paths of Laufer *et al.*'s (2004) subjects reflect the structure of the test they are presented with they are penalized. Hence, if their subjects have not learned the words in the same context settings targeted by the test sentences, they might be unable to recognize that the test sentence is attempting to elicit an item they might know. Alternatively, subjects might know semantically appropriate synonyms (but not the target item) and will not be rewarded for such knowledge. Laufer *et al.*'s (2004) task fails to allow subjects the chance to exhibit knowledge of productive items they may have otherwise. For these reasons, there is no way of knowing whether their test fully reflects the subject's productive knowledge.

Laufer *et al.* imply that productive vocabulary is elicited in response to the prompt provided in their active recall tasks. Yet, and as Chapelle (1998:43) pointed out long before the publication of the test, subjects are required to activate other aspects of knowledge in addition to productive vocabulary knowledge in response to tests, as is the case with the CATSS, such as knowledge of 'context' or 'level of formality' in order to decide which particular words are needed.

2.2.5 Webb (2005): Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge and Webb (2007): The effects of repetition on vocabulary knowledge.

The tasks in the papers reviewed in this section so far each appear to access vocabulary knowledge in a way that elicits multiple aspects of knowledge, in addition to productive vocabulary knowledge. Wesche and Paribakht's Vocabulary Knowledge Scale (for level 5) requires that subjects use words in sentences to confirm knowledge of their 24 pre-selected items, and Laufer *et al.*'s CATSS test requires subjects to respond to sentence completion or multiple choice tasks of pre-selected items and demonstrate knowledge of the contexts provided. Webb also tests with pre-selected items, but is able to confirm lack of prior knowledge because he tests his subjects with nonsense words. His use of nonsense words is important because it makes it possible to see whether subjects' understanding of surrounding items is essential to respond accurately to particular tasks. Webb (2005) describes two experiments, the first of which he tests nonsense words learned from three sentences. He then compares the findings from this first test with a second sentence production task in order to compare gains in receptive and productive knowledge. In his first experiment, Webb (2005) aims to compare learning through receptive tasks with learning through productive tasks. His Japanese L1 subjects were divided into two groups: a receptive group and a productive group. The 'receptive' group was presented with ten target words, along with their L1 meaning, in three sample sentences, as in the following example with 'boulder' (Japanese '巨石') being replaced with 'dangy' (the subjects did not see the Japanese translation): Dangy 巨石

The dangy was as large as a small house.

On the way up the mountain, we passed a dangy.

He stood on the biggest dangy to get a better view.

The 'productive' group met the L2 word and L1 equivalent followed by a space provided to produce a sentence as in the following example:

Dangy 巨石 _____

Both the receptive and the productive group were told to learn the words. The latter were also told to write the words in English sentences. Both groups were given 12 minutes for ten words to complete their learning (and write them, as was the case for the productive group). They were told that upon completion they would be tested, but they did not know how they were to be tested. Each group of subjects was then given a test booklet with one test item on each of its ten pages. The ten target words were presented as nonsense words as in the dangy (boulder) example. For Webb, the use of nonsense words removed the need for a pre-test, as subjects had never met these words before. Webb selected six nouns and four verbs for his ten target words from sentences taken from the 6601 – 14700 most frequent COBUILD words. The nonsense words were: *masco* (locomotive), *denent* (visage), *hodet* (lane), *sagod* (abode), *dangy* (boulder), *tasper* (crave), *cader* (doze), *pacon* (sob), *sagod* (abhor), and *ancon* (dagger). One should note that 'doze' or 'sob' could be interpreted as nouns or verbs. Webb reports that the subjects were tested for knowledge of orthography, association, syntax, grammar, meaning and form. Webb argued that time limitations meant that successful scores on his tests might not indicate that the subjects had acquired full lexical knowledge. Webb notes that full knowledge of the words would involve much more than having read three sentences or writing one.

In Webb's experiment, both groups were tested on both receptive and productive tasks. He assessed his subjects according to the following criteria. In the orthography test, subjects were credited for accurate spelling of a target word they heard twice. Credit for the corresponding receptive test was given if subjects were able to correctly identify target words from a choice of four items (e.g. *dengie, dengy, dungie, dangy*). The productive knowledge of association test offered choices (e.g. *train, airplane,* and *vehicle* to correctly identify a stimulus *masco,* locomotive). The corresponding receptive

task required subjects to isolate the correct associate from a choice of four (i.e. for '*dangy*' from *stone, plant, tree,* and *person*). For orthography, syntax, association, and grammar the receptive tests were in multiple-choice format. Webb's receptive test of meaning used a translation format.

Webb's results indicated that there were gains in all aspects of receptive and productive word knowledge by both groups. However, he notes that the gains were greater for the receptive group than the productive group, and that this runs counter to previous word pair studies (such as Waring's 1997 study) and suggests that the length of time provided for his tasks may have accounted for this. Therefore, Webb conducted a further experiment to determine the precise nature of the gains. This second experiment matched the first, except for four differences: one group took both tasks, the time was limited to the time taken for all subjects to complete, subjects were not told they would be tested, and finally, the groups were tested with 20 target items.

For the second experiment (Webb 2005) an additional ten nonsense words were provided (set B). As in the first experiment, 10 corresponding nonsense equivalents were provided for the target L2 terms. Webb assessed a different group of subjects for his second experiment. An important difference between the first and second experiment was the time restriction imposed for the first task. For the first experiment, the receptive and productive groups were given 12 minutes to complete the initial task. However, in this second experiment subjects were told to start the second task as soon as the first was completed. The results from the second experiment produced lower scores than the first. For Webb the second experiment failed to determine whether the gains in the first experiment were due to the lack of the time restriction but he noted that having to learn 20 items, as opposed to 10, may have skewed his results. In short, he concluded the results from his second experiment show that productive-based learning tasks promote productive learning and enhance receptive knowledge. Overall, Webb reports that both

the receptive and productive tasks contributed to all ten of the measured aspects of vocabulary knowledge.

In his second paper, Webb (2007) attempted to determine the number of repetitions necessary for successful acquisition of lexical items. As in his 2005 paper Webb used 10 nonsense words. Webb randomly organised his 121 Japanese L1 subjects into five groups which included one control group. Each group was presented with a reading task of ten different sample sentences, each containing a different target nonsense word. The experimental groups were divided according to the number of times each target word was encountered (1, 3, 7, or 10 times). Subjects were allowed four minutes to view each set of sample sentences. After the respective exposures to the items, each group was presented with a test measuring different aspects of vocabulary knowledge. The control group, however, did not encounter the target words, yet were asked to complete Webb's productive knowledge of orthography test. Subjects were allowed four minutes to view each context. Webb selected the contexts based on the following three factors: the total number of words, the frequencies of the words in the sentence, and the anticipated ease of topic comprehension. Contexts were then rated according to the ease with which the meaning of a word might be guessed. Contexts were presented in order of difficulty. In this way, the easiest context was presented first and the most difficult last (or tenth, for instance, for the '10' encounters group).

Each word was met in one or two sentences averaging 14 words. The contexts were selected from a sample of graded readers as in the example for 'ancon':

He was not ill, and of course the beds in the ancon are for ill people. One day in 1884, I saw a picture in the window of a shop near the ancon. As soon as he could walk, he left the ancon and started looking for a ship to take him back to England. I did not talk to him very much at the ancon. I looked at his head and arms and legs and body very carefully. (Webb 2007: 53).

Following this exposure, the subjects were then presented with a series of multiple-choice tests with no time restrictions. Each test item was measured in ten different ways as in Webb (2005), meaning that there were two tests for receptive and productive knowledge of: orthographic form, grammatical functions, syntax, association, meaning and form. In addition, each measurement was tested according to two different sensitivities: partial and full knowledge of each item. By way of an example, Webb's testing included productive and receptive knowledge of orthography. The productive version of this test required subjects to write the target word within ten seconds of having heard it twice. The receptive version required subjects to identify the target word in a multiple choice of four. Webb tests for other aspects of word knowledge including receptive knowledge of meaning and form, productive knowledge of grammatical functions, and so on.

Webb reports that all aspects of the groups' vocabulary knowledge improved with increased exposure. Thus, the group with the highest number of exposures (10) achieved higher vocabulary scores than those groups with fewer exposures (7, 3, and then 1, respectively). For full vocabulary knowledge gains Webb argues that 10 meetings, or exposures, are necessary.

Webb's study, in particular his testing methodology, raises two challenging issues. First, it is worth looking at Webb's use of nonsense words, which he claimed guaranteed no prior knowledge (of the test items). As appears to be the case with the other measures reviewed in this section, the subsequent and claimed acquisition of target items may not have occurred in isolation. For this reason, it is difficult to discount subjects' inferencing skills going to work on the surrounding items in Webb's sample sentences, as subjects might have had to make guesses about the meanings of the surrounding words, as well as the target items. Despite the fact that he controls for frequency, Webb assumes that subjects knew the surrounding words in the test sentences yet there is no confirmation that this was entirely the case. Nation (2005) suggests that in order to understand new items in any given context learners need to understand 98% or more of the context

provided. Yet without a pre-test, there is no way of knowing whether Webb's subjects understood as much as 98%. As we cannot confirm that subjects understood all of the surrounding words, we cannot dismiss the possibility that the subjects success (or lack of it) in the test was due to attempts to understand the surrounding items in the test sentences, as well as the target items.

The second issue, though more general, relates only to Webb's papers in the sense that subjects may have known the substituted target words. Webb assumes that the use of nonsense words guarantees no prior knowledge. For Webb's test of 'productive knowledge of grammatical functions' his subjects are asked to respond by writing the target words within a grammatically accurate sentence. In Webb's sentence 'Many doctors and nurses work at 'ancons' he attempts to elicit the substituted or nonsense word for 'hospitals' from his subjects. A potential problem with eliciting words in this way is that there is no real way of knowing whether prior knowledge of each item, in this case 'hospitals,' might have interfered with the subjects' ability to respond. Subjects who produce the lexical item 'hospital' for this particular task would presumably be penalized for failing to produce 'ancon' and therefore fail to gain credit for knowledge they had otherwise. For instance, once a learner immediately links 'ancon' to 'hospital' it is impossible to be certain whether this is a test of their knowledge of 'ancon' or 'hospital', whichever word they produce.

The potential confusion surrounding Webb's subjects' prior knowledge may have influenced his results both in terms of the inferencing skills necessary to complete the sentences and the potential prior knowledge of the substituted target items. The use of nonsense words is nevertheless an important development, since Webb allows us to question the extent to which subjects have no prior knowledge in testing. Webb's use of sentences presents us with a potential issue of subjects inferring meaning of the target words via surrounding items. What remains unclear throughout both of Webb's studies is the extent to which his subjects understood the items surrounding his target words. The apparent success of testees in Webb's studies demands scrutiny in this sense. It is

possible that subjects with full knowledge of the surrounding words achieved the gains Webb claims, given increasingly multiple exposures to those same surrounding items. The repeated exposures to the contexts may have led to an increased understanding of the surrounding words if, on the other hand, subjects did not know them. Thus, there is no guarantee that the subjects understood all the given contexts and that the inferencing skills necessary to guess the meaning of the required target items were not used for the surrounding items as well. Accordingly, it is highly likely that those subjects given 10 exposures to the test items would outscore those with fewer exposures. On first meeting a sentence, I might not know the target item required. However, after multiple exposures, I might be able to develop an understanding of the context provided after accessing semantic or collocational knowledge. Thus, Webb's test does not exclusively elicit productive vocabulary knowledge, because of the multiple aspects of knowledge elicited by such a task.

Webb (2005, 2007) provided five tasks to demonstrate knowledge of 'nonsense' productive vocabulary items (for orthography, meaning and form, grammatical functions, syntax, and association). Each word was met in one or two sentences, the total of which averaged 14 words. Following a different number of exposures (1, 3, 7, or 10) to each sentence, the subjects were then presented with a series of multiple-choice tests with (taken from the first 2005 experiment) and without time restrictions (2007). In addition, each measurement was assessed according to two different sensitivities: partial and full knowledge of each item in his 2007 paper. The productive version of his orthographic test required subjects to write the target word within ten seconds of having heard it twice. Webb suggests that the greater number of repeated exposures (to productive vocabulary items) led to subjects more accurately reproducing knowledge of productive items. The numerous elements in his testing imply that Webb perceives productive vocabulary knowledge as being multifaceted.

2.2.6 Laufer and Nation (1995): Vocabulary size and use: Lexical richness in L2 written production.

Webb's testing with nonsense words is important since it shows that subjects appear to access other aspects of knowledge in addition to productive vocabulary knowledge. Accordingly, the sentence completion tasks reviewed so far, in which subjects are required to understand the surrounding items in the test sentences as well as the part of speech of the target item, all show that productive vocabulary is not being accessed discretely. Laufer and Nation's (1995) paper, compared to the earlier studies in this section is revealing as it provides an alternative method of eliciting productive items with a composition task.

Laufer and Nation (1995) describe a way of measuring the proportion of advanced words in L2 learners' texts. They compare their Lexical Frequency Profile (the FA test in Laufer and Paribakht's (1998) study reviewed in 2.2.3 above) with an 'independent' test (their active version of the levels test (Productive Levels Test) (Laufer and Nation 1995)).

In order to generate a Lexical Frequency Profile (LFP), software compares the proportions of words from different frequency bands in learner compositions, without lemmatization, against vocabulary frequency lists (Laufer and Nation 1990). Laufer and Nation provide an example (table 2.3) of the profile of an intermediate learner's 200-word composition.

Table 2.3 Example profile of intermediate learner (Laufer and Nation 1995:312)

Frequency band	1000	2000	UWL	NiL
Words produced	150	20	20	10
Percentage of words produced	75%	10%	10%	5%

For their experiment, Laufer and Nation divided tthree groups according to proficiency and each completed the Productive Levels Test amd the composition tasks within the same week. The subjects responded to two generaal composition questions, for each of which they had one hour. Laufer and Nation then checked their subjects' compositions by correcting misspellings and removing incorrectly used items. Compositions of at least 300 word tokens in length were then processed using the software described above. Laufer and Nation's three proficiency groups' (from 1 to 3, of which group 3 was the highest proficiency) mean LFP percentages were calculated in order to determine whether the LFP scores indicated both different language proficiency levels and correlations with the Productive Levels Test. The resultant profiles are shown in table 2.4 below.

Table 2.4 Mean percentages and standard deviations at different frequencies (Laufer and Nation 1995:316).

	First 1000		Second 2000		UWL		Not in Lists	
Group	Compl (sd)	Comp2 (sd)	Compl (sd)	Comp2 (sd)	Compl (sd)	Comp2 (sd)	Compl (sd)	Compl (sd)
1	86.5% (3.8)	87.5% (5.3)	7.1% (2.0)	7 % (2.0)	3.2% ((1.8)	4.1% (2.5)	3.3% (2.3)	2.8% (1.8)
2	79.7% (5.3)	79.4% (4.5)	6.7% (1.7)	6.8% (2.2)	8.1% ((2.3)	7.8% (2.3)	5.6% (3.5)	6.6% (3.3)
3	77.0% (6.1)	74.0% (5.9)	6.6% (2.6)	5.6% (2.5)	8.1% ((3.2)	10 % (2.9)	7.5% (2.9)	8.7% (3.5)

In order to compare their LFP profiles with a Productive Levels Test score, Laufer and Nation needed to generate an LFP 'score', which was calculated by tallying the percentage of infrequent words produced (2K+ UWL+NiL). With reference to the different use of frequency words, Laufer and Nation claim that the respective increase or

decrease of rarer word usage for the higher and lower proficiency groups demonstrates evidence of validity of the LFP. Thus, they claim, the LFP was able to distinguish between higher and lower proficiency learners. In order to demonstrate concurrent validity the results were then compared with the Productive Levels Test. Laufer and Nation found high correlations between the levels test and the LFP. In particular, they found that learners who scored well in the Levels test used rarer words in the composition task. Laufer and Nation found no significant differences between the two composition tasks for the lower proficiency learners. They claim, therefore, that the LFP is stable except for their more proficient learners who demonstrate greater variety across different writing tasks.

Laufer and Nation argue the LFP is valid, as it distinguishes between different levels of proficiency. Given the significant correlations with the Productive Levels Test, they also claim that the LFP demonstrates concurrent validity. Based on the assumption that the two tests are broadly operating in the same areas, Laufer and Nation report that the LFP "can reasonably expect learners' vocabulary size as measured by a vocabulary test to be reflected in the learners' productive use" (p. 319).

Laufer and Nation's assumption needs examining in detail since the demands called upon by the two tests are quite different not least since the test burden for the Productive Levels Test is far smaller for the LFP. The Productive Levels Test provides significant clues to test takers compared to the LFP. The Productive Levels Test requires an understanding at the level of the sentence in order to complete a word (e.g. at the 2000 word level, 'Each living room has its own pri____ bath and WC'). Yet the LFP requires compositional skills beyond the level of a sentence completion task such as the Productive Levels Test. There are other potential weaknesses with the Productive Levels Test, see 2.2.2. For this reason, comparisons with the LFP appear to be a hit and miss affair, particular when one considers that for the LFP any words produced within the frequency bands are acceptable, rather than the 18 pre-selected items for the Productive Levels Test.

Having discussed the reasons why Laufer and Nation's tests may not necessarily predictive of one another, we need to explore whether the LFP composition task might inform our investigation of productive vocabulary. A fundamental concern with the LFP composition task is that it imposes significant constraints on subjects. In order to provide a comprehensible composition for the LFP task, subjects need to produce structure for their texts. These structures require a significant number of potentially constraining function words. As Laufer and Nation's (1995) results suggest, a vast proportion of subjects' compositions are made up of such function words and the LFP calls upon an understanding of wider context in a way the other tests described in this section do not. Once the necessity of function words is taken into account, scores for the remaining frequency bands (the 2K+UWL+NiL bands) leave little opportunity for subjects to gain credit when producing infrequent items (with scoring (non 1k) items (i.e. any items produced that fall within the 2k+UWL+NiL bands), typically amounting to 13-23% of Laufer and Nation's subjects' texts). The LFP task requires subjects to produce approximately 300 words on a given subject, which, if Laufer and Nation's data is applied as a standard, leaves only a maximum of 23% or approximately 70 tokens with which to produce infrequent items. Subjects therefore are constrained by the composition task when responding to the LFP task and this imposes restrictions on their ability to display their productive vocabulary knowledge.

Laufer and Nation (1995) wanted to test their subjects' abilities to produce vocabulary within a 'free' composition task. However, their 'free' composition task requires that subjects respond to the contextual constraints of the task, which implies that subjects' responses to a composition task are therefore constrained in terms of lexical content by the need to provide structure to their texts. Thus, the vast number of structure words required by the LFP appears to limit the extent to which subjects can display their productive vocabulary abilities. Thus, the greater the contextual demands the task demands, the further away we appear to be when attempting to assess our subjects' productive vocabulary abilities. Alternatively, we might therefore logically claim the

opposite in which the less a group of subjects is constrained, the closer our access might be to their productive vocabulary knowledge.

Laufer and Nation's treatment of the composition text implies that correct spelling and appropriate use demonstrate knowledge of productive vocabulary. Their LFP credits correctly spelled items if they are produced within a comprehensible context and if they respond appropriately to the two composition questions Laufer and Nation provide in their task. The LFP task reflects Chapelle's (1998: 43) perspective that vocabulary knowledge is inseparable from other aspects of linguistic knowledge since subjects have to consider factors such as sentence or paragraph context and level of formality, as well as needing to know which particular words are necessary. In short, the LFP suggests vocabulary knowledge is reflected by knowledge of appropriate use and accurate spelling. A more detailed analysis of the LFP might show the extent to which subjects typically respond to such tests as the debate in the following two papers (Meara 2005, Laufer 2005) highlights.

2.2.7 Meara (2005): Lexical frequency profiles: A Monte Carlo Analysis and Laufer (2005): Lexical frequency profiles: From Monte Carlo to the real world. A response to Meara.

Meara's (2005) paper evaluates the Lexical Frequency Profile in the form of computerised modelling. Meara says a thorough evaluation of the LFP is necessary because of claims that the LFP:

- "discriminates between learners at different proficiency levels"
- "correlates with an independent measure of vocabulary knowledge"
- "is a useful diagnostic test"
- "is a sensitive research tool" (2005:33)

He adds that there are still several problems with the task such as uncertainties about the way texts are processed (such as whether to ignore or correct errors), or the potential difficulty of having to gather sufficiently large groups of subjects who might typically be reluctant to take part in testing. Meara claims that an additional difficulty is that working with Laufer and Nation's existing data might prove problematic. In order to overcome these two issues Meara proposes modelling subject data in the form of computer simulations.

Meara's model is based on Laufer and Nation' assumption that subjects with access to larger vocabularies typically generate texts that mirror their vocabulary. Meara's 'Monte Carlo' analyses generate simulated texts based on a law that states that the more frequent an item is the lower it is ranked on a frequency table (Zipf 1935). A weighting is included so that the more frequent a word is, the more likely it will be selected while the less frequent words are, the less likely they will be selected. Table 2.5, below, shows Meara's transformations of the numbers from 1000 to 10000. In order to generate a text representative of a subject with a 2000 word lexicon, Meara first calculates Ln (Logarithmic weighting) (2000) x 1000 giving 7600. In order to generate a simulated text of 300 items, random numbers between 1 and 7600 are generated 300 times. If the random numbers produced are less than 6907, then a 1000 word is added to the simulated text. If the random numbers produced are between 6907 and 7600, then a 2000 word is added to the simulated text. The result, for Meara, is a text where the majority of words are 1k words (90%) for a subject with a 2000 word lexicon. Meara shows that his simulations are 'not wildly dissimilar' (2005: 38) to Laufer and Nation's (1995) original data, and with his analysis, Meara claims the ability to control the texts in order to evaluate the LFP.

Table 2.5 Values of Ln(x) x 1000 for a range of values of x Meara (2005: 36).

1000	6907
2000	7600
3000	8006
4000	8294
5000	8517
6000	8699
7000	8853
8000	8987
9000	9104
10,000	9201

In his evaluation, Meara claims that his simulations show only partial support for Laufer and Nation's (1995) earlier claims. Meara concludes that where there was minimal variation between his simulated subject data, he found support for Laufer and Nation's (1995) subject data. Yet where the variation between his simulated subject data was not minimal, Meara suggests Laufer and Nation's claims of stability are 'overgenerous.'

Laufer's (2005) response to Meara (2005) is important since she argues that subjects' responses to test types may not follow a uniform pattern. Laufer first highlights that a Lexical Frequency Profile reflects the proportion of 'use' in writing and not 'size' and argues that Meara is wrong to assume that all areas of lexical competence develop in broadly similar ways. Where Meara suggests the LFP is poor at discriminating between 'groups that are evenly matched' (p.40) Laufer argues that this difference is probably due to learner proficiency and not a weakness in the LFP measure. Her argument suggests that there is potential variation in the way testees respond to different measures either in terms of test type or supra-test considerations (such as motivation, L1 background, and so on).

Laufer contends that size and use probably do not increase in similar ways. Her assertion has broad implications for the LFP and other measures. Laufer suggests that Meara's 'various misinterpretations' (2005: 582) come from his incorrect assumption that size and use do indeed increase in similar ways. For his part, Meara bases his modelling on this general assumption that subjects with larger vocabularies typically generate texts that reflect their vocabulary size. Yet it is evident from Laufer's arguments that this may not always be the case, and not only because of subject abilities. This ability to provide an index of the 'richness' of a text relies not only on the size of a subject's vocabulary but also on individual subject performances. These performances may relate to a changing number of considerations, such as a response to the genre of a question or an anticipation of the audience reading the text. Laufer's contention is that there is not necessarily a relationship between size, and the vocabulary richness she claims the LFP measures. The assumption that subjects with greater access to lexically rich or advanced items produce more lexically rich compositions may not necessarily always hold true.

Laufer's assertion that size and use probably do not increase in similar ways is intriguing. Laufer claims that lexically rich compositions may not necessarily only come from advanced subjects or from subjects with access to such lexically rich items. This is an important development since it might shape the way future tests are approached. Up until this point we have assumed that subjects' vocabulary and use were related in such a way that a test eliciting one might provide an indication of the other. This is important since it may mean that we should look for separate tests to access each aspect. In addition, Laufer's assertion may also mean that we should not necessarily consider that all tests elicit subjects' productive vocabulary consistently, for all subjects, or that we should not expect all subjects to respond to productive measures in the same or similar ways.

For Laufer, then, considering data from computer simulations might cloud rather than clarify the issue. Meara models his data from Laufer and Nation's (1995) original LFP data and we have no real way of knowing if this is reliable in any way. Meara defends

the use of Monte Carlo simulations by claiming that studies involving genuine subjects and that are unsupportive of the LFP are unlikely to have been published, adding that earlier studies are unlikely to have been published if the results were insignificant. Yet Meara then goes on to state that the generated simulated texts are reflective of non-native speaker usage and 'more or less closely' match (p.37) Laufer and Nation's (1995) original test data. Meara adds that as the data broadly mirror Laufer and Nation's original data 'we can use the simulations to test whether LFP really works' (p.38). One might assume that subject behaviour is somewhat more chaotic than the algorithms Meara applies to test the LFP, as Laufer also suggests and for Meara to base his simulations on Laufer and Nation's subject data adds to the confusion, particularly since Laufer argues his data are 'too different' (p.582) from the original LFP data.

Meara suggests that the use of modelling is an 'empirical question that needs further study' (p.46), yet drawing firm conclusions from such modelling might prove difficult given the potential inconsistencies of subject test performances. Without recourse to supportive data, this is a somewhat catch twenty-two conundrum, as such modelling might always be unacceptable unless supported by authentic subject data. Without recourse to supportive and genuine or real data across a range of abilities and performances, research might be no further forward than before Meara's paper.

The LFP appears to constrain subjects in different ways than the sentence completion tasks reviewed earlier in this section, but it appears similar to these other tasks because subjects are required to demonstrate knowledge of multiple aspects of knowledge. The discussions of these tasks in this section suggest that researchers should aim to identify tasks that minimize any potential influence from other types of lexical knowledge when attempting to elicit productive vocabulary knowledge. The Lex30 studies reviewed below, in which I aim to look at papers that introduce, report, and expand on the use of Lex30, might provide us with such a task.

2.3 The Lex30 studies.

This second section presents the background to Lex30. The authors of the Lex30 test claimed their task overcame some of earlier tests' discrepancies and complications, some of which were raised in the first section of the literature review. I will examine what Lex30 considers productive vocabulary to be and consider how successfully it measures it. A review of the Lex30 studies is therefore included in order to draw comparisons with the studies reviewed in section one. Such comparisons might begin to shed light on the extent to which Lex30 exclusively accesses productive vocabulary where earlier tests may not. The first of the five Lex30 studies is Meara and Fitzpatrick's (2000) pilot study and the remaining four papers deal broadly with how well Lex30 elicits productive vocabulary when compared to some of the tests reviewed in section one. In this section, I aim to explore the assumptions Lex30 makes about the construct of productive vocabulary.

2.3.1 Meara and Fitzpatrick (2000): Lex30: An improved method of assessing productive vocabulary in an L2.

Meara and Fitzpatrick's pilot study signalled a departure from other vocabulary measures and addressed, they claimed, a complex 'chicken and egg' situation in the field of vocabulary research. This, they suggested, arose from the lack of well-established and easy-to-use tests of productive lexical skills.

For Meara and Fitzpatrick (2000), the Productive Levels Test (Laufer and Paribakht 1998), reviewed in section 2.2.3, does not exclusively test productive vocabulary knowledge. In addition, the Lexical Frequency Profile (Laufer and Nation 1995) (LFP), reviewed above in section 2.2.6, appears far from efficient at eliciting a representative sample of a subject's ability to produce infrequent items Such tests are either, Meara and Fitzpatrick claim (2000: 21-22), too context specific or fail to provide sufficient information from which to extrapolate learners' vocabulary abilities. Another major

concern is the time taken to complete such tests. In order to address these supposed weaknesses, Meara and Fitzpatrick (2000) designed their Lex30 test.

Meara and Fitzpatrick's (2000) pilot study productive vocabulary test using a word association task, Lex30, presented non-native speaker subjects with a list of 30 stimulus words in English, the L2. As in the example shown below, subjects were required to produce up to four L2 words in response to each stimulus word:

Table 2.6 Lex30 example task.

Look at the words below. Next to each word, write down any other words that it makes you think of. Write down as many as you can (more than 3, if possible). It doesn't matter if the connections between the word and your words are not obvious; simply write down words as you think of them.

CUE	RESPONSES
1. attack	
2. board	
3. close	
4. cloth	
5. dig	
6. dirty	
7. disease	
8. experience	
9. fruit	
10. furniture	
11. habit	
12. hold	
13. hope	
14. kick	
15. map	
16. obey	
17. pot	
18. potato	
19. real	
20. rest	
21. rice	
22. science	
23. seat	
24. spell	
25. substance	
26. stupid	
27. television	
28. tooth	
29. trade	
30. window	

Meara and Fitzpatrick's (2000: 22) criteria were that they wanted cues that prompted a high proportion of varied and infrequent responses. Meara and Fitzpatrick's stimulus words were therefore selected and based on the following three criteria, to:

i. Minimise the influence of receptive vocabulary knowledge on the scores
ii. Differentiate as much as possible between test takers
iii. Allow subjects as much opportunity as possible to produce infrequent vocabulary items.

All of Meara and Fitzpatrick's stimulus words came from Nation's (1984) 1k word list. The stimulus words were selected on the basis that they did not usually elicit a single, dominant response. Hence, words that exceeded more than 25% of native speaker equivalent reported responses (as reported in the Edinburgh Associative Thesaurus, Kiss *et al.* 1973) were rejected. Meara and Fitzpatrick (2000) wanted to avoid a typically narrow set of responses or typically common responses. Then potential cue words were scrutinized for the responses they elicited. If over 50% of the most common responses were not included in Nation's (1984) 1k word list, the item was included in the Lex30 list. Subjects' responses were individually processed, and a mark was given for each infrequent vocabulary item a subject produced. Meara and Fitzpatrick assign infrequent words as those that fall outside of Nation's (1984) first 1000-frequency band. In the pilot study, Lex30 scores represent the total number of infrequent words a subject produces. In later Lex30 studies (Fitzpatrick and Meara 2004, Fitzpatrick 2007), individual scores were calculated by adding up all of the infrequent words that were then scored as a percentage of the total number of words produced by each subject, to minimise the influence of the subjects producing variable numbers of responses.

Lex30 responses were allocated according to the level in which they fell. Accordingly, Level 0 are high frequency words, proper names and numbers; Level 1 words are from the first 1000-frequency band; Level 2 words are made up of the second-1000 frequency band; Level 3 and beyond words are made up of words that fell outside of the first two

frequency bands. A subject's Lex30 score is assessed by counting the number of words produced within the Level 2 or Level 3 and beyond bands for which each item scores one point, or, in short, any response not in level 0 or 1 scores a point. The maximum a subject can score on the test is 120, but this is only theoretical. Table 2.7 illustrates a typical results profile from their subject group (this subject scores 50).

Table 2.7 Typical profile generated by Lex30 (Meara and Fitzpatrick 2000: 24).

	Level 0	Level 1	Level 2	Level 3+
Subject 1	4	49	10	40

The Lex30 pilot study consisted of a comparison of Lex30 scores with scores from another vocabulary test, the Eurocentres Vocabulary size test, or EVST (Meara and Jones 1990). The EVST tests knowledge of words up to a ceiling of 10,000 words from the Thorndike and Lorge (1944) list and the final test score is representative of the subject's vocabulary size. The EVST tests receptive knowledge of vocabulary in the form of yes/no questions as to whether subjects "know" a given word, assuming a direct relationship between the likelihood that a subjects will know the word and its frequency The test begins with the easiest (or most frequent words) and gets progressively more difficult (with less frequent words). Hence, the test starts with the first thousand words and proceeds in sequence up to the tenth thousand words. At each 1000-word band, the program tests subjects with a random sample of 20 words. The EVST stops once it finds a low enough level of performance and carries out a detailed analysis at that level. The test includes a third of 'non-words' and subjects score according to the number of correctly identified words or 'hits' compared to incorrectly identified non-words or 'misses'. Consequently, subjects who carefully identify only words they really do know are credited compared to those who incorrectly identify non-words as 'known'. Accordingly, subjects who incorrectly identify non-words as real English words have their scores reduced.

Meara and Fitzpatrick's subjects were 46 adult learners of English, from a variety of L1 backgrounds, with language proficiency ranging from high elementary to advanced level. Lex30 and the EVST were completed within the same week. Meara and Fitzpatrick found a high correlation between the two sets of scores (0.841 ($p<.01$)), indicating that the two sets of scores were largely predictive of one another. Meara and Fitzpatrick interpreted the results as showing that subjects who produced a large number of infrequent words in the Lex30 test tended to demonstrate a large receptive knowledge of vocabulary as indicated by the EVST test.

Meara and Fitzpatrick claimed that Lex30 is not context specific and is comparatively efficient. Later work by Fitzpatrick and Meara (2004), Fitzpatrick (2007), and Jiménez Catalán and Moreno Espinosa (2005) went on to test Meara and Fitzpatrick's (2000) claims. Baba (2002) preceded these studies with her call for a test of the reliability and validity of Lex30 in her critique reviewed below. In contrast to the studies reported in section 2.2, Meara and Fitzpatrick offer methodology which appears to enable productive vocabulary to be accessed with minimal influence from multiple aspects of knowledge. Fitzpatrick (2007) suggests that Lex30 accesses productive knowledge via only a semantic stimulus (p. 53) and is different from other tests (which have more stimuli) such as the Productive Levels Test which, for example, has three: semantic, orthographic, and collocational stimuli (Fitzpatrick 2007: 52). Lex30 requires that subjects respond by providing associations to the given stimuli, and only illegible associations are removed. If we are determined to measure productive vocabulary, then we need to do so in a way that requires minimal access to other aspects of L2 knowledge. The Lex30 test, certainly when compared to the measures reviewed in section 2.2, appears to offer potential in this respect. However, we still need some way of ensuring that the scores subjects achieve reflect their productive vocabulary knowledge since Meara and Fitzpatrick (2000) only compared their subjects' scores with a receptive vocabulary test. The review of Baba (2002), that follows, suggests how examination of Lex30 ought to proceed.

2.3.2 Baba (2002): Test review: Lex30.

Baba (2002) assesses the potential of Lex30 as a practical testing measure and comments that Meara and Fitzpatrick (2000) fail to adequately demonstrate test validity or reliability. To establish validity, she suggests that further studies need to demonstrate that Lex30 is sensitive to learners' vocabulary improvement over a period of forced practice, and when Lex30 scores are compared with those from other tests of productive knowledge, such as Laufer and Nation's (1999) Productive Levels Test, we need to determine whether subjects achieve similar test scores. To establish reliability, Baba suggests a test-retest approach or a parallel forms approach to measure the internal consistency of the test.

Baba raises two further potential issues with Lex30. She first suggests that the written form might prove "irrelevantly difficult" (p. 70) for L2 learners who might not be able to recognise or respond to the written stimulus words correctly (but might otherwise speak English well). Second, Baba argues that a significant limitation is the apparent inability to tally relative gains as a result of class teaching intervention to any vocabulary produced in the test.

Baba concludes that Lex30 may have a strong impact on L2 vocabulary research as it "exclusively assesses productive vocabulary knowledge" (2002: 70) and goes on to propose a series of measures to investigate the same construct. Fitzpatrick (2007) comments on the available alternative measures in her paper reviewed in the second part of the following section. Fitzpatrick and Meara's (2004) paper, reviewed first below, responds to Baba's concerns. They investigate validity and reliability respectively in one of the two ways Baba proposes.

2.3.3 Fitzpatrick and Meara (2004): Exploring the validity of a test of productive vocabulary and Fitzpatrick (2007): Productive vocabulary tests and the search for concurrent validity.

Fitzpatrick and Meara (2004) and Fitzpatrick (2007) respond to Baba's (2002) concerns. Fitzpatrick and Meara (2004) is the first paper to respond to Baba's (2002) suggestions, reporting on three experiments that address the anticipated reliability and validity issues with Lex30. Fitzpatrick (2007) examines the validity of Lex30. Fitzpatrick and Meara test reliability by assessing the same group of subjects with Lex30 under the same test conditions on two different occasions. The experiment tests subjects using Lex30 twice with a three-day gap between the two test times. The period was chosen to allow sufficient time for subjects to forget the stimulus words, as well as their own responses, and to minimize any influence from L2 improvement or attrition. A comparison of scores at the two test times gave a t-value of $t=1.58$ ($p=.135$) indicating that there was no significant difference between the two sets of scores. The correlation (0.866 ($p<.01$)) shows that subjects achieved broadly similar scores at the two test times. The high correlations show that subjects produce similar proportions of infrequent words at each test time and, as such, Fitzpatrick and Meara claim this indicates reliability. After close analysis of the subject corpora, Fitzpatrick and Meara found that around half of the words produced at test time one were also produced at test time two. Despite such reproduction, Fitzpatrick and Meara assert that their results demonstrate Lex30 produced a representative sample of the subjects' overall lexicon, suggesting that Lex30 has a high degree of test-retest reliability.

To demonstrate test validity, Fitzpatrick and Meara examined the performance of two different subject groups on the same test. In order to demonstrate credible validity they chose to identify a test group known to have a certain level of ability in the trait, productive vocabulary knowledge. They decided to evaluate how well non-native speaker subjects performed with the Lex30 task compared with 46 adult L1 speakers of English from Britain and North America. They compared the native speaker results with

the results obtained from their original (2000) pilot study, which used the same number of subjects. Their results show that the native speakers tended to score higher than the non-native subjects (an average Lex30 score of 44 compared with 30, respectively). Results from a t-test show that the native speakers scored consistently higher than the non-native speakers ($t= 7.5$ $p<.0001$). Fitzpatrick and Meara found that there was a distinct difference between the two sets of average scores and that there was a substantial degree of overlap between the different subject groups. They conclude that their "study demonstrates that the Lex30 test has some validity. Insofar as the design of the test allows, it can distinguish native speakers from non-native speakers" (2004: 66).

In order to test concurrent validity Fitzpatrick and Meara wanted to explore agreement between Lex30 and other tests for productive vocabulary knowledge, (the pilot test only compared Lex30 with a test of receptive knowledge, the EVST). They examined their subjects with two other measures, the Productive Levels Test (Laufer and Nation 1999) and a straightforward translation task from L1 to L2. The concurrent validity test is described in detail later in this section in the review of Fitzpatrick (2007). Fitzpatrick and Meara found a strong correlation between the Productive Levels Test and the translation test (0.843 ($p<0.01$)), which they claim was expected given that the results reflected knowledge of the first 3000 words. I intend to explain this in detail in my review of Fitzpatrick (2007) below. Correlations between Lex30 and the Productive Levels Test (0.504 ($p<.0.01$)) and the translation task (0.651 ($p<0.01$)) were not as high. For Fitzpatrick and Meara, this difference arises because Lex30 scores are less dependent on words from the first three thousand bands, than scores generated by the Productive Levels Test and the translation task. Lex30 awards marks for responses outside the first 1000 word list and the majority of those responses are from the third thousand and beyond (Meara and Fitzpatrick 2000). Nevertheless, they report that the three tests are broadly operating in the same area.

Fitzpatrick and Meara claim that their results show that Lex30 is able to distinguish between native and non-native speaker subjects and that, therefore, Lex30 has some

validity. They also suggest that the collateral tests highlight the concern that the construct of 'productive knowledge' is a complex one. They note that that Lex30 fails to provide evidence that subjects know how to use the words they produce in any meaningful way. The assumptions behind Fitzpatrick and Meara's conclusions need examining in detail.

Fitzpatrick and Meara aim to show that by presenting subjects with the same Lex30 stimulus words at the two different test times the profiles of the responses might remain broadly the same. They did indeed show this, but a substantial proportion of the responses were the same, which may suggest their results ought to be treated with some caution since subjects may have achieved similar scores by simply reproducing the same or similar sets of responses. In this respect, we may be no further forward in establishing reliability. A closer examination of subject data is supportive of this argument. The amount of overlap between subjects' responses over test time one and test time two increases consistently with the number of responses per subject. In other words, those subjects with consistently fewer responses had fewer overlaps and those subjects with a consistently higher number of responses over the two test times had more overlaps. While this is statistically inevitable, what is not clear is whether the limited variation in these sets of responses, over the two test times, is due to subjects exhausting their lexical stores in response to the Lex30 stimulus words or whether it is indicative of stronger and more stable lexical connections in the more proficient subjects. We are left needing to show whether the responses from the subjects accurately reflect the subjects' lexical stores. Presenting the subjects with a different set of different stimulus words might demonstrate that Lex30 is reliable.

Fitzpatrick and Meara's second validity test examined correlations between three measures of productive vocabulary ability. For Fitzpatrick and Meara the lack of a high correlation between all three tests indicated that Lex30 is tapping different areas of vocabulary knowledge and suggests that we need to be careful when claiming to assess 'productive vocabulary' in different tests. It is clear that a definition of the constructs

being measured is needed. That Lex30 might not measure vocabulary knowledge as well or as accurately as other tests is a possibility and one that clearly needs exploring in the experimental chapters.

The experiments reported in Fitzpatrick and Meara (2004) may support the reliability and validity of Lex30 but the residual issues described above need exploring. Before viewing Lex30 as a meaningful tool in the measurement of productive vocabulary, we need to test its validity. Further test-retests (with different cues) might support Fitzpatrick and Meara's claims since there is no guarantee that testing simply exhausted their subjects' lexical stores in response to the same stimulus words.

Fitzpatrick (2007) assesses the concurrent validity of Lex30 by comparing data with Laufer and Nation's (1999) Productive Levels Test, and a translation task. Fitzpatrick chooses these 'collateral tests' based on five shared characteristics, all three tests: i) are based on the assumption that a subject's vocabulary can be measured; ii) purportedly test productive vocabulary; iii) are conducted in a similar way, with pen and paper and take between 15 and 60 minutes; iv) rely on comparison of subject data with frequency lists; v) test vocabulary rather than syntactic knowledge. Fitzpatrick (2007) conducted the Productive Levels Test using five frequency bands: the 2000, 3000, and 5000 word levels, the University Word List level, and the 10000 level. Laufer and Nation's original (1999) study was reviewed in section 2.2.2 above. The Translation task required the subjects to translate 60 words from their L1 (Chinese) into English (the L2). The words were selected at random from the first three of Nation's (1984) frequency bands, with 20 words randomly selected from the first 1000, 20 from the 2000 band, and 20 from the 3000 frequency band. The Chinese translation of the target words was provided, along with the first letter of the English target word, to reduce the effects of synonyms and homonyms. Subjects were awarded a point for each correctly produced word. Misspellings were not penalized. Fitzpatrick attempted to test the validity of Lex30 on the assumption that the three tests measure the same 'ability' (Bachman, 1990: 248) of productive vocabulary knowledge. She tested 55 Chinese intermediate to advanced

learners of English all of whom completed all three tests within two class sessions, one day apart. Fitzpatrick's results indicate a good range of scores as table 2.8 shows.

Table 2.8 Lex30, translation, and Productive Levels Test scores. Fitzpatrick (2007:124).

	N	Mean (sd)
Lex30	55	27 (11.99)
Translation Test	55	32 (12.59)
Productive Levels Test	55	17 (10.55)

The correlations in table 2.9 indicate that a subject scoring well in one of the three tests is likely to score well in the other two but Fitzpatrick (2007) urges caution and notes that we cannot compare scores as like for like.

Table 2.9 Correlations between test scores (Fitzpatrick 2007: 124).

	Productive Levels Test	Translation task
Lex30	0.504 ($p < .01$)	0.651 ($p < .01$)
Productive Levels Test		0.843 ($p < 0.1$)

Fitzpatrick contends that the stronger correlation between the translation task and the Productive Levels Test is likely because the two tasks only really elicited knowledge up to the first 3000 word levels. Fitzpatrick reports that, on average, her subjects produced 14 correct answers at the 2000 and 3000 word levels, and only three correct answers at the other three levels. Given that the translation task only elicited knowledge of words taken from the first 3 thousand word levels Fitzpatrick suggests that a strong correlation is likely with the Productive Levels Test. She also highlights another fundamental difference between the two alternative test formats and Lex30 as the latter gives credit for any infrequent item produced whereas the Productive Levels Test and the translation task require demonstration of knowledge of only predetermined words. Fitzpatrick also notes a further difference between the three tasks in terms of level of difficulty. The

Productive Levels Test and translation tasks increase in difficulty as subjects advance because they encounter increasingly infrequent test items. Lex30 does not become more difficult as subjects proceed through the test.

The differences above lead Fitzpatrick to question whether the three tests are addressing the same ability. For Fitzpatrick each test accesses the lexicon in different ways, or, rather has different 'activation properties' (p.127). She refers to Nation's list of aspects of word knowledge (1990: 27), contending that the construct of 'productive vocabulary can be a misleading label' (p. 131). Fitzpatrick argues that research should clarify which aspects are being addressed by which measures.

Fitzpatrick refers to Bachman to point out the danger of 'leading to an endless spiral of concurrent relatedness' (1990:249). The danger, as Fitzpatrick sees it, is that the pursuit of an appropriate measure of validity will become a cyclical one because researchers may endlessly extend the notion of validity by consistently referring to other tests and their related criteria.

Fitzpatrick's paper ends with the suggestion that future researchers should improve the understanding of the construct of productive vocabulary by comparing new tests with existing tests in order to provide greater understanding of whether a new test has a degree of validity. She also suggests that researchers should consider including other factors such as the influence of learner types, the threshold at which receptive knowledge becomes active, the effect L1s have on test performance, and the relationships between word knowledge and the impact on advancing stages of learner development.

Fitzpatrick's (2007) paper raises the important issue of the need to discriminate between different measures. She uses the three tests of productive vocabulary knowledge in her study to argue, persuasively, that different aspects of lexical knowledge are being accessed. Fitzpatrick's paper shows that it is important to separate precisely which tests measure which aspects of vocabulary knowledge. Fitzpatrick's paper suggests that

Lex30 represents a step forward in the search for an exclusive measure of productive vocabulary while warning of the potential danger of constantly cross-referencing in order to claim validity. The questions raised in Fitzpatrick's (2007) paper provide a context for the following section and final two papers in this chapter which both attempt to describe the construct of productive vocabulary.

The second of the three Lex30 studies to follow Baba's (2002) comments, from Jiménez Catalán and Moreno Espinosa (2005), is reviewed below. Although their paper does not deal exclusively with the issues of reliability and validity Baba raises, it highlights difficulties I should examine. It also highlights the fact that other researchers took Lex30 on board and used it in a large-scale study, which increased the need for further validation.

2.3.4 Jiménez Catalán and Moreno Espinosa. (2005): Using Lex30 to measure the L2 productive vocabulary of Spanish primary learners of EFL.

Jiménez Catalán and Moreno Espinosa suggest that the results from the two earlier Lex30 studies (Meara and Fitzpatrick 2000, Fitzpatrick and Meara 2004) are difficult to interpret because of the degree of variation between subjects. The authors hoped to override such a 'shortcoming(s)' by controlling the following variables in their 2005 experiment: L1 background (all their subjects are Spanish native speakers) and age. Jiménez Catalán and Moreno Espinosa conducted their study with Lex30 on a group of 282 10-year-old Spanish primary school learners of EFL.

Jiménez Catalán and Moreno Espinosa wanted first to determine whether their subjects had sufficient English ability to respond to the cue words, so they examined their subjects' English textbooks. They concluded that all the cue words, selected from Nation's 1k word list (1984), used in the study, consisted of basic vocabulary they expected their young learners to know. I should point out that Jiménez Catalán and Moreno Espinosa scored their subjects' responses using the newer version of Lex30 (v. 2.01) replacing Nation's (1984) word lists with the JACET8000 list (Jacet 2003). The

new version of the test aims to represent more accurate scoring (Fitzpatrick and Meara 2004: 71).

Jiménez Catalán and Moreno Espinosa used the same scoring procedure as Meara and Fitzpatrick (2000: 23). They presented a paper version of the Lex30 test in which they asked their subjects to provide up to four responses for each Lex30 cue. They then typed each of their subject's papers into a text file which is read by the Lex30 (v. 2.01) scorer to determine each subject's Lex30 score. The Lex30 v. 2.01 scorer allowed test takers to produce four responses (the earlier version allowed three) and replaced Nation's (1984) word lists with the JACET 8000 list (Jacet 2003). The Lex30 scorer (v.2.01) (http://www.swan.ac.uk/cals/calsres.html) automatically scores subject's text files by reading each text tile and allocating responses into one of four categories (Level 0 words (high frequency words, proper names, and numbers), Level 1 words (the 1000 most frequent English content words), Level 2 words (the 2000 most frequent English content words), the Not in the List (NiL) band (words that are not found in the previous lists). The allocation of scores was conducted in the same way as Meara and Fitzpatrick (2000) reviewed in section 2.3.1. By way of a diversion from Meara and Fitzpatrick's (2000) original, Jiménez Catalán and Moreno Espinosa also decided on the following criteria for scoring:

- Spanish L1 words score zero
- Made-up or invented English words score zero
- Misspellings were not taken into account, and misspelt words were allocated according to their relevant frequency
- Words that have an equivalent form in Spanish and English were treated as Spanish if the surrounding responses were written in Spanish, or English if the reverse is true

- Proper names of countries, if written in English, were included in the Not in the Lists (NiL) band (words that were not found in the Level 0, Level 1, or Level 2 bands)

Jiménez Catalán and Moreno Espinosa compared their Lex30 scores with the receptive vocabulary size version of the Vocabulary Levels Test (Nation 1983). Jiménez Catalán and Moreno Espinosa adapted the Vocabulary Levels Test to include meanings in Spanish in the 1000 word section (e.g. by asking subjects to find *'black'* for the Spanish equivalent of *'negro'*), in the 2000 word section, the meanings were given in English (e.g. by asking subjects to find *'wall'* for *'part of a house,'* or *'horse'* for *'animal with four legs'*). Jiménez Catalán and Moreno Espinosa then compared their subject's Lex30 scores with their subjects' scores on the receptive version of the Vocabulary Levels Test. The correlations between the two tests was significant but small (for the 1000 word list 0.369, $p<0.01$; the 2000 word list 0.293, $p<0.01$) leading Jiménez Catalán and Moreno Espinosa to extrapolate that Lex30 and the Vocabulary Levels Test are broadly predictive of each other.

Jiménez Catalán and Moreno Espinosa's study raises three problems that need examining. First, their results are difficult to interpret because their scores are significantly lower and the scoring procedure is so different from Meara and Fitzpatrick (2000). Second, their young subjects produced very few responses and Jiménez Catalán and Moreno Espinosa scored according to raw and not percentage scores. Third, their subjects were children and the frequency lists they scored with ((JACET8000) and Nation (1984)) are derived from adult language.

Jiménez Catalán and Moreno Espinosa's paper raises two important issues: first, we may need to establish a minimum level of proficiency in order to discriminate between subjects; and second, we need to score subjects according to the word lists that reflect our subjects' backgrounds.

2.3.5 Moreno Espinosa. (2009): Young Learners' L2 Word Association Responses in Two Different Learning Contexts and (2010) Boys' and Girls' L2 Word Associations.

Moreno Espinosa's two papers aim examine word association responses with Lex30. In her first paper, Moreno Espinosa's (2009) subjects were 130 Spanish young learners of English who were divided into two groups: A and B. Group A consisted of 65 female Basque students, for whom English was their L3 after Basque and Spanish. Group B consisted of 65 north-Spanish students for whom English was their L2.

Moreno Espinosa used the same scoring procedure as Meara and Fitzpatrick (2000: 23). They presented a paper version of the Lex30 test, and asked their subjects to provide up to four responses in response to each Lex30 cue. On the basis that group A had received 331 hours more English education than group B, Moreno Espinosa predicted that Group A would achieve higher mean Lex30 scores than Group B. Accordingly, in the discussion of her results, Moreno Espinosa states that "Lex30 scores reveal that group A has a slightly higher productive vocabulary size (15.96%) than group B (14.64%)....by recalling a higher number of infrequent words" (p.99). In short, Moreno Espinosa appears to suggest that we can extrapolate subjects' vocabulary size from their ability to produce infrequent items in response to Lex30.

Moreno Espinosa (2010)

In her second paper, Moreno Espinosa (2010) follows on from the work in her earlier (2009) paper, but is different in two important respects. First, Moreno Espinosa (2010) is a 3 year longitudinal study, and second, her study examines both boys and girls, as opposed to only the female subjects she examined in her earlier paper. In this second paper, Moreno Espinosa's subjects were 225 Spanish young learners of English who were divided into two groups: A, and B. Group A consisted of 124 male students and

group B consisted of 101 female students. Both groups had started learning English from the age of 3 and ranged in age from nine to twelve.

Moreno Espinosa used Lex30 in order to examine word associations despite the fact that "it was no designed as a word association task but as an instrument to identify productive vocabulary size on the basis of word frequency bands" (p. 5-6). Moreno Espinosa found no significant difference between her male and female subjects Lex30 scores over the three year period. Moreno Espinosa's results suggest that her subject's knowledge of infrequent items improved over the three year period. Moreno Espinosa's results show that the subject group's Lex30 scores improved over the three year period by 5.06% (grade 4, Lex30 score 9.51%; grade 5, Lex30 score 12.12%; grade 6, Lex30 score 14.57%).

Moreno Espinosa's papers raise two important issues relating to the Lex30 cues: first, we might want to consider revising the Lex30 cues since, according to Moreno Espinosa, "[r]ather prototypical clusters of responses [were] elicited by cues such as *fruit* and *furniture*" (2009: 103) which implies we might need to evaluate Meara and Fitzpatrick's claim that "none of the stimulus words typically elicits a single, dominant response" (2000:22); and second, we should treat Lex30 data with caution if we are to interpret vocabulary size based on Lex30 data given that, for instance, some of her subjects repeated the Lex30 cues in their responses (e.g. *attack, close, window*), and may have produced scoring items that are potentially more reflective of textbook usage as opposed to knowledge of infrequent items (e.g. one group produced no words in response to the Lex30 cue *trade,* or *obey* in the 2010 paper, or produced *letter, speak, spelling* in response to the Lex30 cue *spell*).

2.4 Defining the construct of productive vocabulary.

The third and final section in this chapter examines two attempts to define the construct of lexical competence. The first, from Henriksen (1999), is important because she highlights the problem of trying to arrive at a measure that exclusively accesses productive vocabulary. The discussion of her paper serves as a useful framework for the tests reviewed in section 2.1, in order to discuss potential reasons behind the tests failing to measure productive vocabulary. The second paper in this final section, from Read (2004), stresses the need to examine depth of vocabulary knowledge in detail and raises the issue of examining productive 'use'. Read's paper helps to discuss whether we can isolate and examine aspects of use in our attempts to measure productive vocabulary.

2.4.1 Henriksen (1999): Three dimensions of vocabulary development.

Henriksen (1999) suggests that productive vocabulary does not exist in isolation and proposes three dimensions to describe 'lexical competence.' The three dimensions she describes allow us to identify which aspects of knowledge might be measurable in attempts to access productive vocabulary.

In general terms, Henriksen claims researchers have attempted to describe the construct of lexical competence too broadly, by referring to too many dimensions, or too specifically, by referring to too few. Meara argues for only two dimensions suggesting that the construct can be summarized under the headings of *size* and *organization,* because, he claims too many dimensions might confuse attempts to define the construct, "simple dimensions of this type seem to me to offer a rather more promising approach to the problems of measuring lexical competence than do the complex models of vocabulary knowledge" (Meara 1996: 50). Henriksen agrees, in part, with Meara but suggests three dimensions to define the construct.

Henriksen's three dimensions consist of: the partial-precise knowledge dimension, of which she cites various studies ranging from recognition tasks (Palmberg, 1989), to precise tasks such as being able to pronounce a word in an interview task or provide

appropriate associations (Read 1988); the depth of knowledge dimension addresses different levels of lexical knowledge (such as Wesche and Paribakht's (1996) Vocabulary Knowledge Scale); the receptive-productive dimension which addresses different aspects of accessibility (such as selecting an appropriate translation for scoring in a recognition task combined with usage of the same item in a production task). A summary of Henriksen's three dimensions follows.

Henriksen suggests that the first two dimensions, partial-precise and depth of knowledge are related in the process of network building along with mapping of meaning onto form, which she describes as 'semantization' (Beheydt 1987: 57). Henriksen describes 'semantization' as the simultaneous processes in which words are added to and develop within the mental lexicon via a process of deepening understanding of items in relation to other, previously stored, and associated vocabulary. She describes the development of the mental lexicon as an ongoing dynamic process involving the addition, resultant accommodation, and reordering of vocabulary items. Henriksen argues that L2 vocabulary acquisition studies have, to date, often ignored this network building aspect in favour of only 'mapping meaning onto form' (1999: 309). The semantization Henriksen refers to, and her emphasis on network building, may help to interpret responses to Lex30. Unlike many of the tests discussed so far, such as the Productive Levels Test, Lex30 does not make any assumptions about the connections subjects might make between words. However, Lex30 does base its testing on the assumption that it activates a network. As Lex30 does not impose an understanding of a network on subjects, and because subjects are free to respond with their own understanding of the connections they perceive between words, we may be able to interpret subjects' responses as reflective of their own networks.

For Henriksen, the ability to organize the lexicon develops with an understanding of and ability to distinguish between word classes (Miller and Fellbaum 1991), and accompanies a mastery of relating and organizing various semantic groupings. She therefore describes the development of the 'partial-precise' and 'depth of knowledge'

dimensions as interrelated as the mastery of individual terms expands. Henriksen proposes a continuum for the third, 'receptive-productive,' dimension, not a dichotomy, in which items become more familiar and more productive. Henriksen argues that for an item to become more productive, rich meaning representations are important. Henriksen discusses the relationships between these three dimensions of lexical competence. She describes the relationship between the first two dimensions as the flux between a learner developing the sense of a word, and simultaneously developing relationships between it and other related items, and the third dimension as a 'continuum' (p. 313) of related access or use. A fundamental question Henriksen leaves unanswered is the nature of the relationship between the first two dimensions and the third dimension, but she suggests that an increase in ease of access along the third dimension continuum might indicate the extent to which an item has developed along the first and second dimension continua.

Henriksen's paper argues that productive vocabulary should not be considered in isolation, which goes against much of what I have argued so far. In short, the greater the number of different aspects of knowledge we access in addition to productive vocabulary knowledge the less we might claim we have accessed productive vocabulary knowledge. Tests, therefore, that fail to isolate productive vocabulary knowledge cannot discount the possibility that other aspects of knowledge influence results. Henriksen suggests a combination of specific and global measures in order to access the construct of lexical competence and this raises the question of how a combination of tests might measure the construct of productive vocabulary. Given the weakness inherent in adopting either extreme in isolation, one would assume that both approaches, global and specific, might inform each other in the pursuit of assessing productive vocabulary. Yet if one approach is considered with the exclusion of the other, we might miss valuable data. Thus, a 'specific' task that requires the ability to demonstrate knowledge of a particular item, might fail to adequately reflect a learner's productive vocabulary ability if the learner, for instance, is unable to retrieve knowledge of a particular lexical item they might know, otherwise, but cannot demonstrate because of the confines of the task, such as when a subject might know a word (as required by the Productive Levels Test)

but not be able to provide sufficient knowledge of it to respond to the particular sentence context. In addition, a task for 'global' competence may not reveal understanding or rather might shield a lack of specific knowledge subjects may otherwise have. Thus, a task that does not check for specific knowledge might mask detailed information about subjects' knowledge. For instance, a global task might hide the fact that subjects are providing L1 knowledge in an L2 task. For example, responses to Lex30 (Meara and Fitzpatrick 2000) might elicit associations more reflective of L1 usage than the L2 (such as the cue '*pot*' eliciting '*ufo*' (a kind of pot noodle in Japan)). By relying on either global or a specific traits view in the pursuit of productive vocabulary, or a combination as Henriksen suggests, there is a great danger of missing essential data. So-called global tests might fail to fully represent subjects' abilities while specific tests might conceal specific knowledge. Henriksen's proposal to advance earlier attempts to describe all aspects of word knowledge such as those put forward by Nation (1990) or Richards (1976) is convincing yet in order to gain a greater understanding of the construct of lexical competence, both 'specific' and 'global' approaches appear necessary.

2.4.2 Read (2004): Plumbing the depths: How should the construct of vocabulary knowledge be defined?

Read's paper is important for two reasons. First, it raises the issue of the need to examine depth in detail. Second, it advances the assessment of 'use' as an indicator of productive ability. Read's categories of depth studies are worth discussing in some detail since he helps us identify which measures might exclusively access the productive vocabulary construct.

Read suggests that the search for a construct of L2 vocabulary knowledge should elaborate the breadth-depth dichotomy (Henriksen's (1999) second dimension). Read suggests that a single term for depth is inadequate and that research should encompass specific definitions relating to which particular vocabulary abilities are assessed. Read notes that all of the depth studies are limited by two constraints: first, studies reflect

vocabulary knowledge at the time tests are administered (are not longitudinal); and, second, are declarative in the sense that they reflect knowledge learners consciously report.

The following is a summary of Read's three types of studies investigating depth of vocabulary knowledge. Read first introduces 'precision of meaning' studies, referring to Wesche and Paribakht's (1993; 1996) Vocabulary Knowledge Scale as an example of the problems with testing in this area as words might conceivably be unknown and definitions memorized. For Read such representation is problematic because a clear distinction ought to be made between knowing a word and having the ability to define it (Anderson and Nagy 1991). Second, Read describes 'comprehensive word knowledge' studies. He summarizes Nation's (2001: 27) three broad categories of word knowledge (below) to show that such categorization complicates matters considerably because testing means we should assess all the various components of word knowledge.

Form: pronunciation, spelling, word parts

Meaning: form-meaning relationship, concept and referents, associations

Use: grammatical functions, collocations, constraints on use (register, frequency, etc.)

Read argues that such categorization complicates matters considerably if we take testing to mean we should assess all these various components of word knowledge. Read suggests that such analysis would be both time consuming and problematic because we lack the adequate measures or tests to evaluate all of the various components that make up word knowledge. Read argues that it is time consuming because such an investigation might only reveal comprehensive knowledge of only a small set of lexical items. To support this argument he refers to Schmitt's (1998) study, which took two hours to elicit five aspects of eleven words. Read suggests that if research directs its aims towards unearthing a general understanding of learners' vocabulary knowledge we might confuse the focus of assessment.

Read's third categorization of depth studies outlines 'network knowledge' which he suggests is depth of vocabulary knowledge conceived of as a lexical network. For Read the two approaches to depth outlined above focus on the accumulation of individual words, whereas 'network knowledge' studies explore the development of sets of words, and their links, in the mental lexicon. Read describes how lexical growth requires an accommodation of acquired words which also need to be restructured in some way with already known words. In this sense, depth is understood as a learner's ability to distinguish and relate semantically linked items. Read describes how the principal method used to investigate such network knowledge has been the word association task and cites a number of studies addressing L2 word association, including Lex30 (Meara and Fitzpatrick 2000). One should point out that Lex30 is not a word association study as such because Lex30 is designed to elicit infrequent items from subjects and the semantic links are not evaluated. Standard word association tasks subjects' responses are analysed in terms of the associations they provide (e.g. tell me all of the words associated with the word '*think*' that you know (*cogitate, cerebrate, evaluate*, etc.)). Read's word associates task (1998) (http://132.208.224.131/tests/associates/), for example, tests subjects' knowledge of the meaning of adjectives.

Read finally discusses the relationship between breadth and depth of lexical knowledge and cites studies (e.g. Vermeer 2001) supporting his understanding that vocabulary growth is realized in the form of network building and that there should, therefore, be no distinction between breadth and depth. Such studies suggest that a test of one aspect consequently provides an understanding of the other. Read adds that such convergence might only take place at certain levels of proficiency and that lower level learners might demonstrate a greater distinction between these two aspects of knowledge. Read notes that reported word association studies reveal these discrepancies, and that other measures should be implemented in order to measure with more certainty the depth of L2 learners' lexical knowledge. In short, then, Read's call for multiple testing agrees with both Henriksen (1999),who argues for a combination of specific and global trait

measures, and Fitzpatrick (2007) who argues for a focus on comparing existing tests with new tests to improve the understanding of the construct of productive vocabulary.

Considering either of Read's criticisms of 'precision of meaning' studies or 'comprehensive word knowledge' studies (which Read suggests complicate matters considerably since they consider too many factors) there is support for some of the weaknesses outlined in section 2.2. Precision of meaning studies focus on too limited a set of vocabulary and in too detailed a way. In such studies as the Productive Levels Test, subjects are penalized if their knowledge does not reflect the precise item being elicited. Comprehensive word knowledge studies fail to tell us very much about how well subjects know the individual vocabulary items being elicited (and this could be a potential weakness of Lex30).

Read supports word association tasks because he claims they are able to determine whether learners can discriminate between breadth and depth. Read suggests that a word association task is an appropriate task for an investigation of productive vocabulary since, he claims, it provides the ability to separate advanced learners from lower level learners in terms of their ability to distinguish between breadth and depth. This, he claims, avoids two important pitfalls: the first of having to describe meaning in the way 'precision of meaning' studies require; and the second of avoiding having to consider too many factors with 'comprehensive word knowledge studies'. Read's comments relate to word association tasks in general, however, and do not specifically relate to Lex30.

The second important issue is Read's assertion that we should consider word usage to address 'not what learners know about a word but what they can do with it' (p.224). Testing for usage is not without difficulties since it implies a need to administer existing isolated, or Henriksen's (1999) 'separate trait' knowledge, tests alongside tests eliciting use. This discussion relates to Laufer and Nation's (1995) Lexical Frequency Profile (reviewed in section 2.2.6) in the sense that it does test vocabulary in use. The issue of

assessing use is an important one, but potentially far too multifaceted to disentangle the many considerations necessary to apply.

Read's paper is important for the investigation of productive vocabulary for two main reasons. First, although having discussed the issue of 'use' we can see that this is clearly an important issue, fraught with experimental hurdles, because there are too many aspects of knowledge to measure and there is no way of knowing whether these aspects are separable or indeed measurable. Second, Read seems to support Laufer's (2005: 582) warning that subjects may not respond consistently across potentially different test types, which is a fundamental concern for any measurement. This second reason relates to the concern that different test types might activate different aspects of knowledge (Fitzpatrick 2007) and potentially measure different abilities that might be 'irrelevant' (Messick 1989:88-89) to the focus of the test.

2.5 Discussion.

We might now be in a better position to see that some of the tests reviewed in section 2.2 are accessing multiple aspects of lexical knowledge. The reviews of the measures in section 2.2 show that the tests elicit other aspects of knowledge as well as productive vocabulary knowledge. Table 2.10 summarizes the papers reviewed in section 2.2 in terms of the construct of productive vocabulary.

Table 2.10 Comparison of tests of productive vocabulary knowledge.

Task	**How productive vocabulary knowledge is demonstrated**	**Additional aspects of word knowledge elicited? (adapted from Nation 2001: 27)**
The Vocabulary Knowledge Scale (VKS) Wesche and Paribakht (1996)	By providing an example sentence to demonstrate knowledge of 24 pre-selected items *I can use this word in a sentence.* *E.g.:*_____	Form: written, word parts Meaning: form and meaning Use: grammatical functions, constraints on use
Productive Levels Test Laufer and Nation (1998)	By completing pre-selected words in 90 prompt sentences (consisting of 18 words at each frequency band tested (2K, 3K, 5K, UWL, 10K)) *E.g. 'The railway con_____ the city with its suburbs'* (to elicit *connects*)	Form: written, word parts Meaning: form and meaning Use: grammatical functions, collocations, constraints on use
The Computer Adaptive Test of Size and Strength (CATSS) Laufer, Elder, Hill, and Congdon (2004).	By responding to prompts provided in active recall tasks E.g. to complete sentences provided: *'Turn into water m_____'* (to elicit *melt)*	Form: written, word parts Meaning: form and meaning, concepts and referents Use: grammatical functions, collocations, constraints on use.

Webb (2005, 2007)	By multiple aspects of productive vocabulary knowledge of 10/20 nonsense words (orthography, meaning and form, grammatical functions, syntax, and association). E.g. productive version of orthographic test requires subjects to accurately spell the target word within ten seconds having heard it twice	Form: written, word parts Meaning: form and meaning, concepts and referents Use: grammatical functions, collocations, constraints on use
The Lexical Frequency Profile Laufer and Nation (1995)	By correct spelling and appropriate use of words in response to 300 word composition task	Form: written, word parts Meaning: form and meaning, concepts and referents Use: grammatical functions, collocations, constraints on use
Lex30 Meara and Fitzpatrick (2000)	By providing infrequent L2 responses to 30 cue words e.g. provide up to four responses to cue *attack*	Form: written Meaning: form and meaning, Concepts and referents, associations

Lex30 relates to this survey in the sense that the additional aspects of knowledge that are elicited (as well as productive vocabulary knowledge) appear to be kept to a minimum. At the beginning of section two, we saw that Baba (2002) concludes that Lex30 appears to test productive vocabulary knowledge exclusively. Lex30 is therefore worth exploring in detail given that it is more likely to measure productive vocabulary with fewer of the influences described in section 2.2 Lex30 appears to overcome many of the issues raised by earlier tests of productive vocabulary. The aim, therefore, should now be to examine Lex30 in detail in order to determine precisely what it is measuring and how useful it might be in terms of accessing vocabulary in a meaningful way. There is support from the papers reviewed in section 2.4 in which both Henriksen (1999) and Read (2004) argue in favour of a particular kind of experimentation necessary to measure the construct of productive vocabulary. The three dimensions Henriksen presents allow us to examine and separate precisely which aspects of these dimensions might be measurable. Read suggests that a word association task appears to avoid the pitfalls of alternative measures, by providing the ability to distinguish between advanced and lower level learners. In terms of these arguments, there is some support to justify further investigation of the construct of productive vocabulary with the Lex30 test.

There are clear variations between the different test types, as sections 2.2 and 2.3 show, and potential variations among testees' responses to different tests as discussed in section 2.4. My questions and experiments in the following chapters are based on the issues I have identified with existing tests of productive vocabulary and unresolved problems with Lex30 as identified by Meara and Fitzpatrick (2000), Fitzpatrick and Meara (2004), Fitzpatrick (2007), Baba (2002), and Jiménez Catalán and Moreno Espinosa (2005).

Lex30 is worth exploring in detail because there are important outstanding issues to explore which relate to its validity and reliability as a test of productive vocabulary. With this exploration in mind, the following presents a summary of the research questions for the five experimental chapters, which follow.

Chapter three replicates Meara and Fitzpatrick's (2000) pilot study and asks whether a replication of this study reproduces the same or similar sets of data with a different group of subjects under the same test conditions. The research question for chapter three is: Can a replication of Meara and Fitzpatrick's (2000) study reproduce the same or similar results with a different group of subjects under the same test conditions?

Chapter four addresses whether subjects with low levels of orthography in the L2 (such as Chinese or Japanese subjects, whose L1 does not use the roman script) may be at a disadvantage with the written form of Lex30 (Baba 2002). The three research questions for chapter four are: a) Is there a significant difference in the way subjects perform on a spoken and written response Lex30 format?; b) How do the correlations between X_Lex and Lex30 in this experiment relate to those in chapter three and in Meara and Fitzpatrick 2000 (which used EVST instead of X_Lex)?; and, c) Is there a threshold number of responses below which Lex30 does not work?

Chapter five examines whether a different, but similarly selected, set of cue words, from different frequencies, might produce similar results to the Lex30 original, and how, if at all, do individual subjects' Lex30 (and other formats of Lex30) scores change over a 6-week test period. The three research questions for chapter five are: a) Will a different, but similarly selected, set of cue words produce similar results to the Lex30 original?; b) Will cue words from a different frequency band (2k) produce different scores?; and, c) How, if at all, do individual subjects' Lex30 (and other versions of Lex30) scores change over a 6-week period?

Chapters six and seven address Fitzpatrick's (2007) suggestion that we compare existing tests of productive vocabulary. Chapter six first compares subjects' Lex30 scores with LFP scores. The research question for the first study in chapter six is: Is the proportion of infrequent items derived from Lex30 different from that derived from a discursive task? A Brainstorm Frequency Profile (BFP) task (introduced in the second study in

chapter six) was devised in order to compare responses to the LFP task and to elicit freely selected items, with the same context but without the need to provide function words or composition structure. For the Brainstorm Frequency Profile, subjects brainstorm their responses to a Lexical Frequency Profile question as opposed to writing a 300-word composition (as is standard for the LFP task). The Brainstorm Frequency Profile task elicits 120 items, the same number of items as Lex30. The research question for study two asked: Is there a closer relationship between Brainstorm Frequency Profile and Lex30 scores than between LFP and Lex30 scores?

Chapter seven compares subject performance on Lex30, a GapFill task, and the Productive Levels Test. The Gap Fill task was devised in order to compare responses to the Productive Levels Test, which is designed to elicit pre-determined items, and to compare some of the features of the Productive Levels Test in order to elicit a sample of a subject's lexicon and to compare those results with Lex30 scores. The aim of the GapFill task is to elicit a range of possible answers in the form of responses to a sentence completion task and to determine whether subjects provide a greater proportion of responses compared to the Productive Levels Test. With the GapFill task, there is a semantic stimulus and syntactic context (as with the Productive Levels Test and Lex30) but there is no form stimulus and so subjects are able to respond with any word, not a single pre-determined item as is required by the Productive Levels Test. The research question for chapter seven is: What is the relationship between GapFill scores and Lex30 scores, compared to a task designed to elicit pre-determined productive vocabulary (The Productive Levels Test)?

Finally, in my discussion chapter, I aim to discuss the findings from the experimental chapters and discuss them in light of the literature review. The concluding chapter collects the main strands of the thesis and proposes possible areas for future research.

2.6 Conclusion.

The literature review examined the extent to which the construct of productive vocabulary is accessed and measured by the tests reviewed. The examination appears to suggest that the majority of tests appear to activate multiple aspects of knowledge. Lex30 appears to activate very few aspects of knowledge and the experimental chapters that follow investigate this claim in detail.

Chapter 3 Replicating Meara and Fitzpatrick (2000)

3.1 Introduction.

In chapter two, I identified a number of potential issues that arise when attempting to measure productive vocabulary. The issues relate to the studies that attempt to elicit productive vocabulary when tests: i) require subjects to define items for which they may lack the relevant vocabulary (they may know the item but not the words required to define the item (e.g. Wesche and Paribakht 1996)); ii) attempt to assess subjects with items that are more infrequent than the vocabulary knowledge being accessed; iii) assume that knowledge of one aspect of lexical knowledge presupposes another, in a straightforward way; iv) use nonsense words when subjects may know the replaced items; and, v) composition writing needing function words, and a focus on compositional structures, which might muddy the assessment of productive vocabulary.

In response to these issues, raised in the first section of chapter two, the second section raised the question of whether we have an appropriate measure of productive vocabulary with Lex30, because with the Lex30 task (and in response to the five issues above) subjects: i) do not have to define target items; ii) are only assessed on their ability to provide any infrequent items in response to the highly frequent cues; iii) are not assumed to have the same kinds of structures in their lexical stores; iv) do not need to provide knowledge of nonsense words; and, v) do not need to respond in composition form.

Considerable further investigation is necessary before establishing Lex30 as an exclusive, efficient, accurate, and valid measure of productive vocabulary and Fitzpatrick (2007) contends that any claims of robustness of the Lex30 measure might be premature, because collateral tests may not measure the same or even overlapping aspects of the same construct. As the review of the tests in the first section in the second chapter suggests, testing for validity is problematic because it is difficult to find other tests that measure the same ability. In an earlier test for concurrent validity, Fitzpatrick and Meara (2004) sought to determine whether Lex30 distinguishes between native and

non-native speakers and found that "there seems to be a good degree of overlap between the scores of the [two] groups" (p.63). Other issues need addressing, validity aside, such as whether Lex30 scores are dependent on the particular cues chosen (or whether cues selected according to the same selection criteria as Meara and Fitzpatrick (2000) elicit similar scores), and the extent to which Lex30 can measure productive vocabulary ability without much of the influence of other aspects of knowledge.

Meara and Fitzpatrick argue that, as discussed in detail in 2.3.1 above, there is a lack of well-established and easy-to-use tests of productive lexical knowledge and refer both to Laufer and Nation's (1995) free productive vocabulary test (reviewed in 2.2.6) and Laufer and Nation's (1999) controlled productive vocabulary test (reviewed in 2.2.3) as examples. For Meara and Fitzpatrick (2000: 19-21) the weakness of such tasks stems from their view that it is difficult to extrapolate information about subjects' lexicons from such small samples of text, and that such tasks are not efficient because of the limited amount of class time usually available. Meara and Fitzpatrick claim Lex30 responds to such weaknesses, suggesting it is worth examining.

In order to begin examining Lex30, I replicate Meara and Fitzpatrick's (2000) pilot study. Meara and Fitzpatrick (2000) found that Lex30 scores correlated well with an independent measure of general proficiency, the Eurocentres Vocabulary Size Test (EVST). If, in the replication, I find that Lex30 scores correlate well with an independent measure then this may justify further experimentation with Lex30 to test productive vocabulary with Lex30. This chapter, therefore, replicates Meara and Fitzpatrick (2000) to investigate the robustness of their findings. Accordingly, the research question for this chapter is:

Can a replication of Meara and Fitzpatrick's (2000) study reproduce the same or similar results with a different group of subjects under the same test conditions?

3.2 The replication study.

Section 2.3.1 presented a detailed summary of Lex30 as reported in Meara and Fitzpatrick (2000). My aim here is to refer to the similarities and differences between their study and the replication reported below.

3.2.1 Subjects.

The replication subjects were a group of 50 university students studying English as a foreign language (EFL), aged between eighteen and nineteen, and made up of thirty-two males and eighteen females. The majority of the subjects were Japanese L1 speakers with a minority with L1 Mandarin, L1 Korean, and L1 Cantonese. The students took three hours of English language classes a week within the university; the classes took the form of speaking practice in which students discuss social issues. The English proficiency of the subject group, which was judged to range from elementary to pre-intermediate (the students had not taken another independent test, such as TOEFL or TOEIC), was not as diverse as Meara and Fitzpatrick's "Arabic to Icelandic" (p.23) subject group, which was a group of 46 EFL learners ranging from high elementary to proficiency level.

3.2.2 Method.

The replication study subjects were given instructions in order to respond in the same way as Meara and Fitzpatrick's (2000) subjects. To this end, the subjects were asked to write up to four response words for each cue provided. The subjects were given an example to begin with. I wrote a word on the classroom blackboard and then elicited four responses from the subjects. Meara and Fitzpatrick (2000) do not report giving an example, but I wanted the students to try to provide as many words as possible and considered that working through an example response might facilitate this. Following

the demonstration, the Lex30 cue words were presented as a list and the test took approximately 15 minutes to complete. For an example of a completed replication test refer to Appendix 1. In the same week as the Lex30 task, the subjects completed a standard yes/no test of receptive knowledge (X_Lex) (Meara and Milton 2002). X_Lex was used in place of the EVST (Meara and Jones 1990) which Meara and Fitzpatrick (2000) used in their original. The more recent X_Lex task was used in this replication because it was considered to be more sensitive to lower proficiency levels, as it tests knowledge up to 5000 not 10,000 words. While the EVST tests from the first ten thousand words, X_Lex targets the first five thousand. Both the EVST and X_Lex tests require subjects to respond to a computerized test by reporting whether they know a given pseudo or genuine word or not (section 2.3.1 describes the EVST in detail). X_Lex is delivered on a computer and the calculation includes an error adjustment for the number of non-words subjects claim to know, like the EVST. A sixth of the presented words in X_Lex are pseudo words or distractor words, and in a similar way to the EVST, subjects' approximate vocabulary size is calculated from the number of correctly identified genuine words or false alarms. The number of false alarms each subject claims to know is taken into account, along with correctly identified genuine words. As a result, those subjects who incorrectly identify false alarms (as genuine words) have their vocabulary scores reduced according to the number of non-words they claim to know, subjects who carefully only identify genuine words do not have their scores reduced.

3.2.3 Scoring.

Each set of Lex30 responses was typed into a computer text file in order to score the data using the Lex scorer software (v. 2.01) (Meara and Fitzpatrick 2004). In the same way as the original study (Meara and Fitzpatrick 2000: 29-30) subject data were lemmatized using Bauer and Nation's word family categories for level 2 and 3 affixes (Appendix 2). By choosing only to lemmatize words with relatively frequent affixes

(level 2 and 3 of Bauer and Nation's lists), credit was given for use of less frequent morphology. Subject number one's response 'virtually' to the 19th Lex30 cue 'real,' for instance, contains a level 3 affix '-ly' and was lemmatized to 'virtual'. Subject number ten produced 'amusement' (NiL) in response to the 27th Lex30 cue 'television,' (the suffix '-ment' attached to the verb 'amuse' to make the noun amusement) was not lemmatized because it is not in Bauer and Nation's level 2 and 3 lists. Each text was processed using the Lex scorer (v. 2.01) (Meara and Fitzpatrick 2004) which produces a frequency profile according to Nation's (1984) frequency list. The scorer profiles subjects' responses according to the number of level 0, level 1, level 2, and level 3 words. Meara and Fitzpatrick (2000) also scored their data according to Nation's word lists (1984). Any obvious misspellings were corrected, and all proper nouns were counted as level 0 words. Each subject's Lex30 score was calculated in the same way as Meara and Fitzpatrick (2000), where any word produced outside of the level 0 and 1 band scored one point (i.e. Lex30 score = Level2 +Level 3+ words). Hence, subject number one's score of 27 (shown in table 3.1) was calculated by tallying the total number of words produced in the level 2 band (15) and the level 3 and above band (12).

Table 3.1 Lex30 score generated by subject 1.

	Level 0	Level 1	Level 2	Level 3 +	Lex30 score
Subject 1	5	30	15	12	27

3.3 Results.

3.3.1 Lex30 scores.

The results appear to indicate that the replication follows a broadly similar pattern to Meara and Fitzpatrick's study. The Lex30 results for the replication and Meara and Fitzpatrick's (2000) original are shown in table 3.2 below, I have included both sets of results in order to compare the Lex30 scores at each level. The most obvious point of

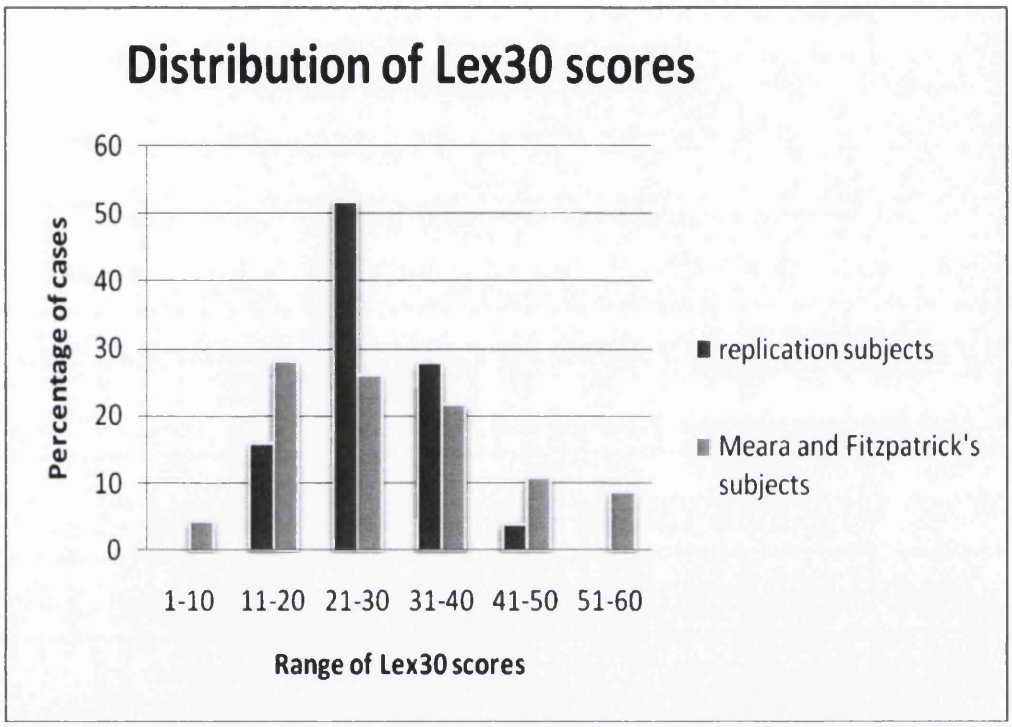
comparison is that the mean scores are very similar. For both studies, most of the words produced by the subjects fall into Nation's (1984) first thousand category. The replication study subjects produced more level 2 than level 3+ responses while Meara and Fitzpatrick's (2000) subjects produced more level 3+ than level 2 responses. The replication subjects produced an average of 60 responses to the 30 cues (Meara and Fitzpatrick's subjects produced an average of 90 different words).

Table 3.2 Lex30 mean score and standard deviations (sd).

	Level 0	Level 1	Level 2	Level 3+	Total words	Lex30 score
Replication	2.21 (2.2)	33.5 (12.4)	15.5 (6.2)	11.9 (2.1)	63.4 (16.9)	27.4 (7.3)
Meara and Fitzpatrick (2000)	3.7 (3.6)	59.3 (13.9)	7.8 (3.6)	20.8 (11.4)	91.6 (24.2)	28.9 (13.9)

Once the raw scores were converted to percentage scores (mean replication Lex30 percentage score: 43%, mean Meara and Fitzpatrick (2000) Lex30 percentage score: 31%) we see two important differences. First, the difference in percentage scores (between the replication and Meara and Fitzpatrick's (2000) study) highlights the difference in number of scoring responses. Second, the difference in raw and percentage scores highlights the importance of having a percentage score as well as or instead of a raw score. Figure 3.1 below shows the different distribution of Lex30 percentage scores for the two studies and that the scores are clearly distributed differently. The cases are represented as percentages for clarity as both tested a different number of subjects (Meara and Fitzpatrick tested 46 subjects while the current study tested 50 subjects). The replication subjects' scores are clustered in the 21-30 range, while Meara and Fitzpatrick's subjects' scores are less clustered and more evenly distributed between the 11-40 range.

Figure 3.1. A comparison of Meara and Fitzpatrick's (2000) subject data with the replication as percentages.



3.3.2 Comparisons with yes/no test.

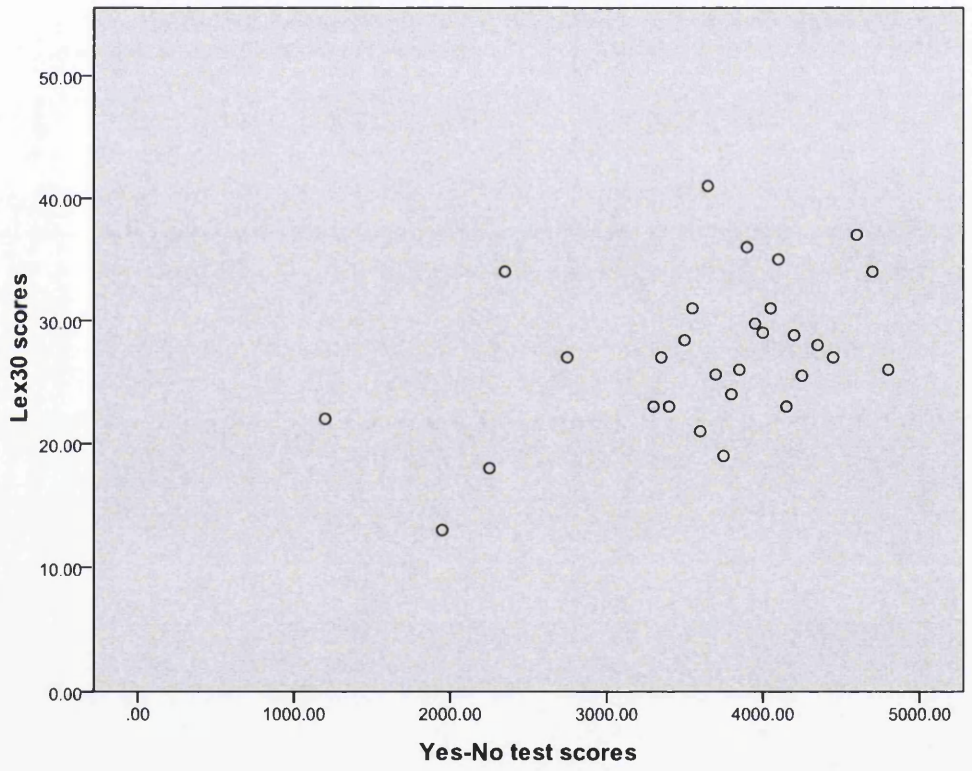
The yes/no test scores, shown in table 3.3 below, indicate that the lower scoring replication study subjects were more tightly clustered in terms of their receptive knowledge compared to Meara and Fitzpatrick's higher scoring subjects. The Lex30 scores in this replication were compared with the independent receptive measure X_Lex. The correlation between the two sets of scores in this replication was not significant (0.274 (p =0.54)) while Meara and Fitzpatrick's correlation was significant (0.841 ($p<0.01$)). Meara and Fitzpatrick (2000: 24-5) suggest that their significant correlation indicates that the two tests are largely predictive of each other, while the correlation from this replication does not support their claims. Figure 3.2 shows the relationship

between the Lex30 and the X_Lex scores, and shows that the subjects X_Lex scores are clustered around a score of 4000. Both the clustered X_Lex scores, and the clustering of Lex30 scores around 30, shown in figure 3.1 above, represent a lack of diversity with the replication subject group.

Table 3.3 A comparison of means and standard deviations of yes/no tests.

	Replication (X_Lex) (sd)	Meara and Fitzpatrick (2000) (EVST) (sd)
Mean	3717 (670)	5089 (2089)

Figure 3.2 Comparison of X_Lex and Lex30 scores.



3.4 Discussion.

The main purpose of this study was to test the robustness of Meara and Fitzpatrick's pilot Lex30 study with a replication. The research question asked whether a replication of Meara and Fitzpatrick's (2000) study would reproduce the same or similar results with a different group of subjects under similar test conditions. The Lex30 mean raw scores were comparable (mean replication Lex30 score: 27.4 (sd. 7.34), Meara and Fitzpatrick (2000) mean Lex30 score: 28.9 (sd. 13.9)) but the standard deviations suggest a greater range of scores in the original study than the replication. The correlations between the yes-no test and Lex30 in this replication (0.274 ($p = 0.54$)) do not support Meara and Fitzpatrick's (2000) finding of a strong and significant correlation between Lex30 and an independent receptive measure (the EVST) (0.841 ($p<0.01$)), which obviously needs explaining. In this discussion, I aim to investigate two issues: first, I examine the difference between the subject groups, and the different receptive measures. Second, I look at two issues that arise from the replication: the quality of word knowledge, and the different numbers of words produced.

The first issue relates to the potential homogeneity of the replication subjects, in terms of their L2 proficiency, when compared to Meara and Fitzpatrick's (2000) subjects, which we see from the raw Lex30 and yes/no test scores. The lower standard deviations for the replication raw Lex30 scores appear to support this claim (replication mean Lex30 raw score: 27.4 (sd 7.34) when compared to Meara and Fitzpatrick's subject group (28.9 (sd 13.9)). The lower standard deviations for the yes/no test also appear to support this claim (replication mean X_Lex score: 3717 (sd 670) compared to Meara and Fitzpatrick's subject group (EVST score: 5089 (sd 2089)). Figure 3.2 shows that although the X_Lex scores ranged from 1100 to 4900 most replication subjects (44) are clustered in the 3500 to 4500 range and as these X_Lex scores represent frequency bands suggests that most of the replication study subjects knew around 4000 words. Without comparing the two measures for receptive ability, it is unreasonable to make general claims about the

difference in (yes/no) test scores and this leads me to this second concern, the difference between the two receptive measures. Even though the EVST and X_Lex are designed according to comparable principles, they are different in one important way. The X_Lex test ceiling is lower than the EVST. X_Lex was used because the subjects' proficiencies appeared considerably lower, on average, than Meara and Fitzpatrick's subjects, but the results suggest that X_Lex was not sensitive enough to measure the vocabulary size of some of the more proficient replication subjects. The X_Lex test has a maximum score of 5000 and only tests words up to the first 5000, while the EVST has a maximum score of 10, 000. X_Lex tests subjects' knowledge with a representative set of words from each of the 5 frequency bands and then estimates each subject's overall knowledge of each frequency band, to produce a total score out of a maximum of 5000. The EVST, by contrast, can investigate word knowledge above the 5000 frequency band to differentiate between subjects who know words, whereas X_Lex does not, above the 5000 frequency band. Bachman suggests that "a ceiling effect might occur when the test is too easy for a particular group of test takers, so that many of them obtain perfect scores" (2004: 96). Given that most of the replication subjects' X_Lex scores lay between 3500 and 4500, X_Lex appears to fail to distinguish between subject vocabulary knowledge as effectively as the EVST. This appears to be due to its comparatively low maximum score, and the fact that some of the subjects score close to "perfect" scores (Bachman 2004:96), which suggests, again, that X_Lex was not sensitive enough to differentiate between the subjects. As most of the subjects scored between 3500 and 4500 on the X_Lex task it appears likely that they would also know some of the words in the 5K+ frequencies.

The second issue relates to two concerns that arise from this study. Two broad observations arising from the data relate to the quality or degree of word knowledge elicited by Lex30. The first concern relates to the way subjects wrote their responses. Five subjects wrote between 15 and 20 of their responses in Katakana (Japanese syllabic writing for words of foreign origin) and so we do not know whether these subjects can represent such responses in the English alphabet. One way of examining this might be to

conduct a spoken response version of Lex30 in order to minimise or eliminate the effect of orthographic ability as Baba (2002) suggests (2.3.2). To determine whether a written test might limit subjects' responses to only producing items they are confident of spelling, and whether written responses to Lex30 might be unrepresentative of subjects' lexicons at the lowest threshold the study reported in chapter four examines whether written and spoken responses elicit different Lex30 scores. The second concern relates to the quality of the associations subjects provide. While Lex30 was not designed to look at the quality of associations at all, it does tell us that the words subjects provide in response to the task are those that exist in the subjects' lexicons. An examination of some of the subjects' responses does perhaps suggest that their knowledge of the words they produced might be incomplete. In response to the cue *map*, for instance, several subjects responded with what appear to be Japanese loan words, such as *fan* and *club,* which one subject explained was because "..it's from a Japanese website". The issue here is not that subjects are producing what appear to be Japanese-like associations (the subjects appear to know that what they are producing are English words) but whether the samples Lex30 elicits are representative of the subjects' productive lexicons. Accordingly, we need some means of determining whether the samples Lex30 elicits are representative of the subjects' productive lexicons. Therefore, to address this second issue, the study reported in chapter five examines subjects with an alternate set of Lex30 cues selected according to the same criteria as Meara and Fitzpatrick's (2000) original Lex30 cues.

One further issue relates to the different number of words subjects produced in response to the Lex30 cues. Lex30 requires subjects to respond with up to four words for each of the thirty cues, a maximum of 120 words. The replication study produced a mean of 63.4 (sd 16.9) words while Meara and Fitzpatrick's (2000) subjects produced a mean of 91.6 (sd 24.2) words. If we examine the Lex30 scores in terms of raw scoring, we see that the replication subject's raw scores (mean 27.4 (sd 7.34)) are very similar to Meara and Fitzpatrick's (2000) subjects' raw scores (mean 28.9 (13.9)) but this way of scoring only appears to tell us the number of infrequent items produced. If, however, we look at the

number of infrequent items produced as a proportion of the total number of words each subject produces, the subjects' percentage scores, we see that the replication subjects achieve higher Lex30% scores than Meara and Fitzpatrick's subjects (mean replication Lex30 percentage score: 43%, mean Meara and Fitzpatrick (2000) Lex30 percentage score: 31%). The fact that Meara and Fitzpatrick's subjects produce a greater number of words overall appears to indicate a relationship between the number of words produced and their L2 proficiency. The numbers of words produced, therefore, appears to indicate other elements of proficiency such as fluency and, conceivably, motivation. Consequently, the greater number of words produced by Meara and Fitzpatrick's (2000) subjects appears to indicate a greater overall proficiency despite the fact that the replication subjects' percentage scores are higher. This potentially higher proficiency is difficult to prove, and one alternative explanation for the different numbers of words produced might relate to the cultural traits of the predominantly Japanese L1 subjects who, potentially, when unsure of a response might have chosen not to respond. The fewer words produced in total by the replication subjects might relate to a threshold number of responses below which Lex30 might become invalid as we discussed in the review of Jiménez Catalán and Moreno Espinosa (2005) (section 2.3.4).

3.5 Conclusion.

The findings from this replication do not appear to support Meara and Fitzpatrick's (2000) with the lack of significant correlation between Lex30 and the independent receptive measure. I discussed two possible reasons for this. The first related to the difference in subject group, and the second related to the difference in receptive measure. The study raised three additional issues for further investigation, namely: the potential influence of orthographic knowledge, the quality of word knowledge, and the concern that there might be a threshold number of responses below which Lex30 might become invalid.

Chapter 4 Comparing written and spoken responses

4.1 Introduction.

Chapter three raised three important issues responding to Meara and Fitzpatrick's (2000) findings. The discussion in section 3.3 first questioned whether subjects knew how to represent their responses in English orthography and suggested a comparison of spoken response Lex30 scores with written response Lex30 scores. Second, in contrast to Meara and Fitzpatrick (2000), the replication study found no significant correlation between Lex30 and the yes-no test. One possible explanation for this different result, discussed in section 3.3, was the issue of the clustered replication subject group. Third, the replication study elicited a smaller mean number of words (63.4 (sd 16.9)) than Meara and Fitzpatrick's study (mean 91.6 (sd 24.2)), raising the issue of a possible threshold level existing below which Lex30 does not work. Each of these three issues is explored in greater detail below.

First, as Baba (2002) conjectures, and as discussed in 2.3.2 and 3.3, the standard written format of Lex30 may penalise subjects for a lack of orthographic knowledge. The discussion section in chapter three (3.3) suggested that subjects' responses to the Lex30 task might only have been the ones that they knew how to write. Baba's argument is that, although they might be able to verbalise responses, subjects may lack the orthographic knowledge to write them. This concern might be especially pertinent for subject groups, such as those tested in chapter three, whose L1 has a different orthography (i.e. Japanese) from English. To explore this, this chapter reports on an experiment that compares Lex30 written response scores and Lex30 spoken response scores, with both sets of responses elicited by using the same Lex30 cues. A lack of significant difference between the results from the two Lex30 formats might indicate, contrary to Baba's conjecture, that a subject's productive vocabulary knowledge can be tapped

commensurately by either a spoken or a written test, with the items that the subjects produce not being influenced by a lack of orthographic knowledge.

Second, and in contrast to Meara and Fitzpatrick's (2000) study, one of the main findings from chapter three was the lack of significant correlation between Lex30 and the independent measure of receptive vocabulary in the replication. One possible explanation for this lack of correlation, discussed in 3.3, was that the proficiency of the replication subject group examined might not have been sufficiently diverse. Another explanation was the different ceiling scores between X_Lex and the EVST. Meara and Fitzpatrick (2000) found a significant correlation between the receptive vocabulary knowledge test and Lex30 (0.841 ($p<0.01$)); however, the correlation from the study presented in chapter three was not significant (0.274 ($p=0.54$)). Meara and Fitzpatrick (2000) observed that their results support suggestions that an increase in receptive vocabulary will be matched by in an increase in productive vocabulary as Laufer (1998: 267)) proposed, for subjects "with vocabularies of 10th and 11th grade learners of English" (Meara and Fitzpatrick 2000: 26). However, the results from chapter three suggest that the relationship may not be quite so straightforward. Thus, one aim of the current study is to determine whether testing with a different, and potentially more diverse, group of subjects might result in a stronger correlation between Lex30 scores and receptive measure scores. If this proves to be the case, we might then claim that the lack of significant correlation reported in chapter three was due to the subject group.

Third, one alternative and plausible explanation for the findings in chapter three is that Lex30 might not be sensitive enough to distinguish between subject groups of similar proficiency. We may discern this possibility in two ways: the number of words produced and the tightly clustered scores. The replication study elicited fewer mean responses (63.4 (sd 16.9)) than Meara and Fitzpatrick (91.6 (sd 24.2)). The standard deviations also appear to suggest that the subjects in the group were too tightly clustered in terms of their ability to respond with infrequent vocabulary items: replication Lex30 mean score (27.4 (sd 7.34)); Meara and Fitzpatrick (2000) Lex30 mean score (28.9 (sd 13.9)). This

lack of significant correlation between Lex30 and the yes-no test reported in chapter three might have been due to a lack of sensitivity in Lex30 when it comes to distinguishing between subjects of similar proficiency. With a more diverse subject group than that examined in chapter three, we might be able to determine whether the lack of significant correlation between Lex30 and the receptive measure is due to the particular subject group of chapter three and not due to any lack of sensitivity with Lex30. If the lack of significant correlations were due to the subject group this might then suggest that there is a level below which Lex30 becomes invalid. This theory is that the lower aggregate number of words produced in total by the replication subjects might indicate a threshold number of responses below which Lex30 might become invalid.

In order to address these three issues, the experiment in chapter four: i) compares spoken response Lex30 scores with written response Lex30 scores; ii) compares Lex30 with a yes-not test with a more diverse group of subjects than those examined in the study reported in chapter three; and iii) examines whether there is a threshold level below which Lex30 does not work. Accordingly, the research questions are as follows:

a. Is there a significant difference in the way subjects perform on a spoken and written response Lex30 format?
b. How do the correlations between X_Lex and Lex30 in this experiment relate to those in chapter three and in Meara and Fitzpatrick (2000) (which used EVST instead of X_Lex)?
c. Is there a threshold number of responses below which Lex30 does not work?

4.2 Study.

4.2.1Subjects.

The subjects were 40 university students enrolled in English as a foreign language (EFL) courses, aged between eighteen and twenty, and made up of twenty-two females and

eighteen males. The L1 majority of the subjects' was Japanese (33 in total), while other subjects had L1 Chinese, L1 Malay, or L1 Korean. The students took one and a half hours of English language instruction a week within the university, which took the form of speaking practice in which students discuss social issues. The English proficiency of the group ranged from high elementary to low intermediate. This group was selected precisely because the L2 proficiency level was moderately more diverse than the subject group examined in chapter three (whose L2 proficiency ranged from elementary to pre-intermediate). I am not able to report test scores from an independent measure, as the students had not taken another independent test, such as TOEFL or TOEIC. The diversity of the group was still not quite as diverse as Meara and Fitzpatrick (2000), in which the subjects' proficiencies ranged from high elementary to proficiency level.

4.2.2 Method.

In the first of the two tests (Lex30written), the subjects were asked to write up to four response words for each cue, in accordance with the standard Lex30 protocol. Subjects were then asked to complete the independent measure of receptive knowledge, the X_Lex task (see section 3.2.2), within the same week. The subjects then had a 6-week gap between taking the Lex30written task and the second task in order to allow sufficient time to forget the cues. Subjects then took the spoken format of the Lex30 task (Lex30spoken). The subjects were given the same written instructions and cues as for the standard Lex30 test with the exception that they were told to verbalise their responses. The subjects read the cue words on individual cards, each card being presented in the same order as the standard written Lex30 task, and they verbalized their Lex30spoken responses with a short (5 second) pause between their final response to each of the cues. Once subjects had produced four words for a cue, I showed them the next cue. If subjects were unable to provide four responses to a cue, they moved on to the next cue. Subjects were asked to point to the cue cards if they were ready to move on to the next cue (i.e. if they could not think of a response). The subjects' spoken responses were recorded on audio tape and later typed into a machine-readable text file.

(Appendix 3 shows a set of sample responses to the Lex30written and Lex30spoken tasks). Both the written and spoken Lex30 task scores were then processed in the same way as by Meara and Fitzpatrick (2000:24) (see section 3.2.2), with the exception that the word lists used (JACET8000) were more recent than the word lists that Meara and Fitzpatrick used to score Lex30 (Nation (1984)). I selected the JACET8000 lists for two reasons: because they are more recent, and because they are designed to be more relevant to Japanese learners (Mizumoto and Takeuchi, 2009: 428-429). Based on the British National Corpus, the JACET8000 ranks English words according to the frequency with which Japanese learners encounter them. The more recent word lists might thus represent a more accurate picture of the subjects' lexicons and be especially relevant for the Japanese L1 subjects examined, because the lists might more closely reflect their learning paths compared to Nation's (1984) word lists.

4.2.3 Results.

The results are given in relation to each research question posed. The first research question asked:

4.2.3.1 Is there a significant difference in the way subjects perform on a spoken and written response Lex30 format?

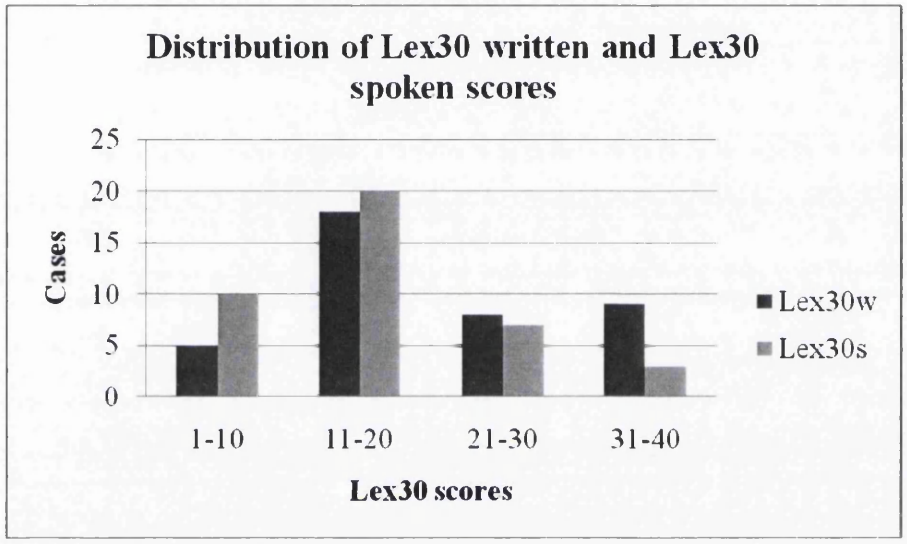
The means and standard deviations, shown in table 4.1, indicate that subjects did not perform very differently on the two versions of the test. Paired t-tests ($t = 0.751$ $p=0.451$) show there was no significant difference between the two sets of scores (Lex30 written and Lex30 spoken).

Table 4.1 Mean scores and standard deviations for Lex30written and Lex30spoken and number of words produced.

	Lex30 mean score (sd)	Number of words (sd)
Lex30written	16.3 (8.1)	47.5 (19)
Lex30spoken	15.6 (7.1)	55.8 (21.6)

Figure 4.1 indicates that responses to both tests follow a broadly similar pattern, although the subjects do perform slightly differently for each version of the task. Slightly more subjects' Lex30 scores seem to be in the high score bands on the written format compared to the spoken format.

Figure 4.1 Distribution of Lex30 written and Lex30 spoken raw scores.



The correlation between the scores for the two versions of Lex30 was moderate but significant (0.568 ($p<.01$)) with a paired t-test showing a significant difference between the number of words produced ($t = 2.76$ ($p<.009$)). As table 4.1 shows, subjects produced a mean of 55.8 (sd 21.6) words in response to the spoken version of the test, and a mean of 47.5 (sd 19) to the written, highlighting that subjects produced fewer items in response to the written Lex30 task than the spoken. Subjects' mean Lex30 written scores (16.3 (sd 8.1)) were higher than their Lex30 spoken scores (15.6 (sd 7.1)). The different number of words produced in response to each test format might have been as a result of the particular test format. The written version required that subjects complete their paper tests within a class group, while the spoken version of the test required subjects to verbalise their responses individually in an audio room. The lower Lex30 spoken scores, despite the greater numbers of words produced, might indicate

that subjects produced more words they felt confident of verbalising. The subjects may have felt inhibited by having to respond to the spoken test format (in front of their native speaker examiner/ teacher) and may have not given themselves time to think of enough responses, possibly explaining the lower scores on the spoken Lex30 task compared to the written task.

4.2.3.2 How do the correlations between X_Lex and Lex30 in this experiment relate to those in chapter three and in Meara and Fitzpatrick 2000 (which used EVST instead of X_Lex)?

Table 4.2 shows that the correlations between Lex30 written and the independent receptive measure are significant (0.530 $p<.01$), though still not as strong as those in Meara and Fitzpatrick's (2000) study (0.841 $p<.01$). These results indicate that a slightly more diverse group produces scores that correlate more strongly.

Table 4.2 Correlations between Lex30 and Yes-No tests in different three Lex30 studies.

Study	Receptive Measure	Correlation
Meara and Fitzpatrick (2000)	EVST (Meara and Jones, 1990)	0.841 $p<.01$
replication (chapter 3)	X_Lex (Meara and Milton, 2003)	0.274 $p=0.54$
replication (chapter 4)	X_Lex (Meara and Milton, 2003)	0.530 $p<.01$
spoken (chapter 4)	X_Lex (Meara and Milton, 2003)	0.429 $p<.01$

4.2.3.3 Is there a threshold number of responses below which Lex30 does not work?

Table 4.3 shows the correlations between Lex30written and the independent receptive measure with different numbers of subjects. This analysis cuts out the very lowest producing (two) subjects to determine if we get a stronger correlation than becomes apparent between X_Lex and Lex30 for the remaining subjects. The lowest producing subject produced only five responses, with the next lowest producing 19 responses, table 4.3 shows the correlations with these subjects removed from the analysis. Removing

these two subjects (who produced the fewest number of words (5 (10+) and 19 (20+)) in response to Lex30, produces moderately stronger correlations, with the strongest and most significant correlation occurring when subjects produce 20 or more words in response to Lex30. Even so, this analysis appears to suggest that there is not a threshold number of responses below which Lex30 does not work.

Table 4.3 Correlations between Lex30 and X_Lex when subjects produce 10, 20, 30, or 40 or more words.

Number of words produced	Correlation with X_Lex
10+	.489**
20+	.541**
30+	.445**
40+	.445**

** Correlation is significant at the 0.01 level (2-tailed).

4.3 Discussion.

The three research questions in this chapter were based on the findings from the replication study in chapter three. The first research question was devised to determine whether there is a significant difference in the way subjects respond to a spoken version of Lex30 compared to a written version. The second question was devised to determine whether an increase in a subject's receptive vocabulary relates to an increase in their productive vocabulary. The third research question was devised to determine whether there is a threshold level number of responses below which Lex30 becomes invalid. This discussion addresses each of these research questions in turn.

The first research question examined whether there is any significant difference in the way subjects responded to a written and a spoken Lex30 task. The first research question explored Baba's suggestion that responses to a written only test might potentially disadvantage those subjects lacking in orthographic knowledge: "one limitation of Lex30 is that it assesses the learners' written performances but does not assess their

spoken vocabulary knowledge" (Baba 2002: 70). However, the Lex30written task generated higher scores (Lex30written mean score: 16.3 (sd 8.1)) than the Lex30spoken task (Lex30spoken mean score: 15.6 (sd 7.1)), while the paired t-test ($t = 0.751$, $p = 0.451$) indicates that there was no significant difference between the two (written and spoken) task scores. On the other hand, the Lex30written task elicited fewer items (Lex30written mean number of responses 47.5 (sd 19)) than the Lex30spoken task (Lex30spoken mean number of responses 55.8 (sd 21.6)) with the paired t-test showing a significant difference between the number of words produced ($t = 2.76$ ($p<.009$)). This significant difference between the numbers of words produced might be due to the subjects feeling inhibited to produce words that they knew, but did not feel confident pronouncing. The lack of formal oral communication classes at Japanese High Schools may have limited the subjects' confidence in the spoken task. Subjects may also have felt inhibited by having to produce and pronounce words in front of their native English-speaking examiner. Alternatively, individual learners might have responded to the formats in different ways. Thus, in response to the first research question, both (spoken and written) Lex30 tasks elicit broadly similar scores, indicating that the Lex30 task seems to tap into both spoken and written productive knowledge in broadly similar ways.

The second research question asked how the correlations between the receptive measure, X_Lex, and Lex30 in this experiment, compared to those in chapter three and in Meara and Fitzpatrick 2000 (which used the EVST rather than X_Lex). A strong correlation, as in Meara and Fitzpatrick (2000), with the receptive measure (X_Lex) would indicate that the two tests are measuring similar aspects of knowledge. Table 4.2 shows the correlations between Lex30 and yes-no tests in the three different Lex30 studies. Based on the non-significant correlation between Lex30 and X_Lex from chapter three, I wanted to test the robustness of the Meara and Fitzpatrick (2000) findings with a potentially more diverse group of subjects in this study. The significant correlation found in this study between X_Lex and Lex30written (0. 530 ($p<0.01$)), therefore, could thus be attributable to the more diverse subject group.

The correlation between Lex30 and X_Lex reported in this chapter is still weaker than that stated in the Meara and Fitzpatrick (2000) pilot study, so I discuss three possible reasons for this below. As I discussed in the introduction (4.1), the relative lack of diversity of the subject group (compared to Meara and Fitzpatrick (2000)) might be a factor. As table 4.1 shows, the subjects produced fewer words in response to both formats of Lex30 (Lex30 written mean words produced: 47.5 (sd 19) and Lex30 spoken mean words produced 55.8 (sd 21.6)), than Meara and Fitzpatrick's subjects (mean words produced 91.6 sd 24.2). Meara and Fitzpatrick (2000) report their Lex30 scores as raw scores, or a straightforward count of the number of infrequent words produced. In a more recent study, Fitzpatrick and Meara (2004) converted their subjects' raw Lex30 scores to percentage scores to reflect the number of infrequent items produced as a proportion of the total number of words produced. In order to determine whether Lex30 percentage scores would show a stronger correlation with Lex30 and the independent receptive measure X_Lex, I converted the Lex30 raw scores from my study to percentage scores, as Fitzpatrick and Meara (2004). However, table 4.4 shows a lack of significant correlation between the Lex30 written percentage scores and X_Lex, and between the Lex30 spoken percentage scores and X_Lex. The lack of significant correlation between X_Lex and either format of Lex30 suggests that the significant variable is the number of words the subjects produced.

Table 4.4 Correlations between Lex30% and X_Lex.

Lex30written % (chapter 4)	0.142 p= 0.381
Lex30spoken % (chapter 4)	0.215 p= 0.183

There are three, other, potential explanations behind the significant, though not strong, correlations between X_Lex and Lex30 written raw scores in this latest experiment. The first explanation relates to the group diversity, as in the replication study, which may explain the weak but significant correlations in the current study. Meara and Fitzpatrick's subjects produced a greater number of words (91.6 (sd 24.2)) than the

subjects in the study reported in this chapter (Lex30 written mean words produced: 47.5 (sd 19), Lex30 spoken mean words produced 55.8 (sd 21.6)). The standard deviations for Meara and Fitzpatrick's Lex30 scores indicate greater subject diversity than the standard deviations for the current study. Meara and Fitzpatrick's subjects' mean Lex30 score was 28.9 (sd 13.9), while the current study's Lex30written mean raw score was 16.3 (sd 8.1) and the Lex30spoken mean raw score was 15.6 (sd 7.1)). The second explanation relates to the different receptive measures used in the current experiment and the replication compared to Meara and Fitzpatrick's (2000) pilot. As discussed in detail in 3.3, the lower X_Lex ceiling score of 5000, compared to 10, 000 for the EVST might explain the non-significant correlation for the study reported in chapter three. The issue here is whether the subjects' vocabulary abilities are reflected more accurately by the X_Lex task than the EVST. For instance, two subjects scored 4700 on the X_Lex task, and their being so close to the 5000 ceiling suggests that these two subjects, and others, too, conceivably, may have performed differently or better had they been stretched more by taking the EVST task (with a higher ceiling of 10,000). The third explanation relates to the scoring for the Lex30 test itself because it has been updated since the original experiment (Meara and Fitzpatrick 2000). Meara and Fitzpatrick compiled their subjects' responses according to Nation's (1984) word lists while this study used the JACET8000 word lists (JACET 2003). Fitzpatrick and Meara (2004) claim that the JACET8000 word lists may provide a more accurate, and up to date, set of scores and this hypothesis might account for the different correlations (compared to the replication). In short, it appears necessary to determine how significantly these three reasons influenced the results. Indeed, the lack of a strong correlation between Lex30written and X_Lex could be explained by one or a combination of these three factors. Further testing is necessary to determine which if any of the three factors (i.e. subjects not being sufficiently diverse, the use of different receptive measures compared to Meara and Fitzpatrick (2000), and the updated scoring) had the greatest impact on the correlations.

The third research question was devised to determine whether there is a threshold level number of responses below which Lex30 becomes invalid. Two subjects produced a

particularly low number of words in response to the written task (5 and 19 words in total). A stronger and more significant correlation between Lex30 (written) and X_Lex might indicate that there is a threshold number of responses below which Lex30 does not work. Table 4.3 shows the reanalyses of the subject data categorised according to the number of words produced (10, 20, 30, and 40+ items) and calculated with their correlations with X_Lex discretely. The correlations between Lex30 and X_Lex for the subjects producing 20 items or more (table 4.4) (0.541 $p<.01$) are slightly more significant than the correlations between Lex30 and X_Lex without removing any of the lower producing subjects (0.530 $p<.01$) from the correlation analysis. Thus, there does not appear to be a threshold number of responses below which Lex30 does not work. However, the reanalysis of the correlations does appear to suggest that when subjects produce 20 or more items in response to Lex30 we get the strongest and most significant correlations between Lex30 and X_Lex.

An additional observation pertains to the question of which level of threshold knowledge is implied by the production of particular words in response to Lex30. In 3.3, we saw that subject responses might not always constitute knowledge of L2 productive vocabulary. Subjects appeared to provide loan word responses to stimuli, such as providing '*royale*' for '*battle*' ('Battle Royale' is a Japanese film), or '*ufo*' to the cue '*pot*' (a variety of Japanese pot noodle). While subjects clearly demonstrate some degree of productive knowledge, as in such cases where subjects produce '*ufo*' in response to the cue '*pot*', it is difficult to ascertain exactly how minimal their understanding of these L2 words might be. Accordingly, such responses to Lex30 merely appear to demonstrate that that subjects are vaguely aware that the items they provide are from the L2 English. For instance then, I am probably far more aware than my Japanese subjects that the adjective 'Royale' is borrowed from the French, and that in contrast to English, French adjectives are usually positioned after the noun they modify. Conversely, though, if subjects are producing loan words in response to the cues (e.g. *royale* to the cue *battle*) this could signify that they are aware that these words are viable in English, possibly

intimating at least minimal knowledge of the item that they provide. There is no huge issue with subjects producing what appear to be Japanese-like associations because it nonetheless seems that the subjects know that they are producing English words; however, what we need to resolve is whether Lex30 elicits truly representative samples from subjects' productive lexicons (i.e. not just automatic pat responses). One means of doing so might be to conduct interviews after the test, although this might prove too time consuming and potentially difficult to validate. In terms of validating an interview, we would need to be careful to ensure that subjects were unhindered by the spoken component of the task. The results in this study appear to suggest that there is a very minor difference between the written and spoken formats of Lex30, in terms of the numbers of infrequent items and the words in total, subjects responded with. We should therefore seek to determine an alternative means of evaluating responses to Lex30, by attempting to eliminate or at least minimise the influence of other factors (such as subjects having the pressure of having to speak in front of the their teacher). An alternative method of evaluating subjects' responses to the Lex30 cues is to evaluate whether the test would elicit similar samples of subjects' productive lexicons with a different set of cues. In other words, the kinds of responses that subjects provided might have been prompted by the particular cues used. Consequently, if we test subjects with a different set of cues and elicit similar Lex30 scores (similar scores to those generated by the original cues), this would indicate that it is indeed Lex30, and not particular cues, that elicits representative samples from subjects' productive lexicons. For these reasons, I aim to examine whether the current Lex30 cues in particular are responsible for the Lex30 scores, or whether any cues chosen according to the same principles would produce the same or similar results. To address this issue, the study reported in chapter five examines subjects with an alternate set of Lex30 cues selected according to the same criteria as Meara and Fitzpatrick's (2000) original Lex30 cues.

4.4 Conclusion.

The first research question examined whether there is a significant difference in the way that subjects responded to a spoken version of Lex30 compared to a written version. An examination of the scores attained on the spoken and the written test formats of Lex30 reveals no significant difference between the two. Both elicited broadly similar scores, so we may claim that Lex30 seems to elicit spoken and written productive vocabulary knowledge in broadly the same way. Thus, Lex30 appears to tap subjects' productive vocabulary knowledge regardless of the task format (spoken or written).

The second question investigated whether an increase in a subject's receptive vocabulary corresponds to an increase in their productive vocabulary. The significant correlations between the Lex30 test and the independent receptive measure (X_Lex) appear to support Meara and Fitzpatrick's (2000: 25-26) observation that the number of infrequent vocabulary items that Lex30 elicits is broadly reflective of the number of items known receptively by each subject, and as measured by an independent receptive measure.

The third research question investigated whether there is a threshold level number of responses below which Lex30 becomes invalid. The reanalysis of the data suggests that, when subjects produce more than 20 responses, the strongest and most significant correlations are manifest between Lex30 and X_Lex; however, the correlations are nevertheless still strong and significant even when subjects are included who produce as few as five words in response to the task. Consequently, this study indicates that no threshold level number of words appears to exist beneath which Lex30 becomes invalid.

By comparing two versions of the Lex30 task, a written and a spoken, I attempted to address Baba's (2002) concerns, and the findings from this second experimental chapter tentatively justify continuing my investigations with the written format of Lex30. The written format of Lex30 correlates with the receptive measure, and with the spoken format. The significant correlations were still not as strong as in Meara and Fitzpatrick's pilot (2000) study, suggesting that I need to explore Lex30 further. I discussed three

potential reasons for the lack of comparably strong correlations, namely: i) the relatively homogeneous proficiency of the subjects compared to Meara and Fitzpatrick (2000); ii) the different receptive measures (with different ceiling scores); and, iii) the updated tools of measurement since the pilot (I used the JACET8000 (Jacet 2003) instead of Nation's (1984) word lists). I also discussed the issue of whether responses to Lex30 are representative of subjects' productive lexicons and postulated the need for a means to establish whether the responses provided by subjects are representative of their lexicons, rather than being merely attributable to the specific original Lex30 cues. In order to address this final concern, the study reported in chapter five examines subjects with alternate sets of Lex30 cues, selected according to the same criteria as Meara and Fitzpatrick's (2000) original Lex30 cues.

Chapter 5 Testing construct validity (with alternative sets of cues) and reliability (over a six-week test-retest period)

5.1 Introduction.

The results from the experiment reported in chapter four revealed no significant difference between the spoken and the written formats of Lex30. The experiment, however, raised the issue of whether responses to Lex30 are representative of the subject's L2 productive knowledge or whether the responses Lex30 generates might be peculiar to the particular Lex30 cues. These included responses such as *ufo* to the cue *pot* or *royale* in response to the cue *battle*. The discussion in section 4.3 suggested two means of determining whether responses to Lex30 were representative of a subject's L2 lexicon. The first of these means was in the form of post hoc interview tests, which was dismissed because we might not be able to examine subjects with the limited amount of time we have with them and because such testing might prove theoretically difficult to validate. The second idea suggested testing the current Lex30 cues, selected according to the same selection criteria as Meara and Fitzpatrick (2000), to determine whether they are responsible for the Lex30 scores, or whether any cues chosen according to the same principle will produce the same or similar results. This second idea implies that testing with an alternative set of cues, that produce the same or similar results, might indicate that Lex30 elicits scores that are representative of each subject's productive vocabulary (regardless of whether subjects appear to produce what appear to be L1 responses to one particular set of Lex30 cues). Accordingly, I aim to examine whether the original Lex30 cues in particular are responsible for the Lex30 scores, or whether any cues chosen according to the same principle produce the same or similar results.

The first aim of this chapter, then, is to vary the cues in order to test Lex30, because, as Bachman says, "the test user may wish to measure individuals' language abilities frequently over a period of time, and want[s] to be sure that any changes in performance are not due to practice effect, and therefore uses alternate forms" (1990, p. 183). The study reported in this chapter selects different sets of prompt words, therefore, according

to the same selection criteria as Meara and Fitzpatrick (2000: 22-23) in order to test with 'alternate forms' of Lex30. If the test scores relate appropriately to a specific generalization (Bachman 1990: 189), in this case the similarity of the cue word selection process, then this might contribute to Lex30's construct validity. In other words, if subjects achieve similar scores from two different sets of Lex30 cues, selected according to the same criteria as Meara and Fitzpatrick, then this will support claims of construct validity of Lex30 as a test of productive vocabulary. We might then be able to claim that Lex30, regardless of which cues we use, is successful at eliciting similar scores that reflect subjects' productive vocabulary knowledge. For this first aim, I selected an alternative set of cue words from the first thousand frequency band (JC1k).

The second aim of this chapter is to explore the effect the frequency of the cue words has on Lex30 scores. I therefore selected a second set of cues from the second thousand frequency band (JC2k). Meara and Fitzpatrick (2000) claim that the original Lex30 cues, selected from the first thousand frequency band, are "words that even a fairly low-level learner would be expected to recognize [and] typically generate(s) responses which are not common words" (2000: 22-23). Accordingly, I aim to test this assumption by comparing scores generated by responses to three different sets of cues from two different frequency bands: Lexorig (Meara and Fitzpatrick's original cues), JC1k (an alternative set of cue words from the first thousand frequency band), and JC2k (an alternative set of cue words from the second thousand frequency band). Meara (1983: 29) suggests that high frequency words tend to produce high frequency responses and a comparison of the three sets of cues investigates whether the lower frequency cues (JC2k) elicit a greater number of lower frequency responses from the subjects' lexicons than the higher frequency cues (Lexorig, and JC1k).

The third aim for this chapter is to test whether Lex30 might be considered reliable. Once we are able to establish the reliability of Lex30 we can then investigate its validity, because "in order for a test score to be valid, it must be reliable" (Bachman 1990:160). While reliability is concerned with "factors other than the language ability we want to

measure" (Bachman 1990: 160-161) Bachman notes that validity is concerned with the extent to which "an individual's test performance is due to the language abilities we want to measure" (1990: 161). The test for reliability examines subjects Lex30 scores at two different test times. If subjects' scores have not changed significantly at the second test time, then this may support the notion that the Lex30 task is reliable (Bachman 1990:160-1).

The aims of the experiment in this chapter are to: i) vary the cues in order to test Lex30; ii) explore the effect the frequency of the cue words has on the Lex30 scores; and, iii) examine subjects' Lex30 scores at two different test times to test for reliability. Accordingly, the three research questions are as follows:

a. Will a different, but selected according to the same criteria, set of cue words produce similar results to the Lex30 original?
b. Will cue words from a different frequency band (2k) produce different scores?
c. How, if at all, do individual subjects' Lex30 (and other versions of Lex30) scores change over a 6-week period?

5.2 Study.

5.2.1 Selection of alternative cue words.

I needed to generate a different set of cues from the first thousand (for JC1k) and second thousand frequency band (for JC2k). The selection process was based on Meara and Fitzpatrick's (2000: 22) criteria to produce cues that have the potential to prompt a high proportion of varied and infrequent responses. Meara and Fitzpatrick's stimulus words were selected based on three criteria: first, the words were taken from Nation's first thousand-word list (1984) so that they would be recognized by most proficiency levels of L2 learners; second, they selected words that tended to elicit infrequent responses. To do this they looked up potential stimulus words in the Edinburgh Associative Thesaurus (EAT) (Kiss *et al.*, 1973), and rejected stimulus words that

elicited predictable (e.g. *black> white*) responses. Meara and Fitzpatrick rejected any primary responses that exceeded 25% of the reported EAT responses. This second criterion was devised to elicit a range or a great variety of responses to maximize chances of differentiating between subjects. Third, they wanted to elicit infrequent items, so cues were included if at least half of the most common EAT responses were not included in the first thousand word lists. With the aim of providing a more recent and potentially more appropriate set of cues for the Japanese subject group, the alternative cues (JC1k and JC2k) were chosen from the JACET8000 word lists (Jacet 2003).

In short, for the selection process, the alternative cue words were selected if they met the following three criteria:

i. All JC1k cue words were selected from the first thousand JACET8000 word list. By selecting cues from the first thousand frequency band, in their case from Nation's (1984) word lists, Meara and Fitzpatrick describe the importance of including this to "make it possible to use the test with learners across a wide range of proficiencies" (2000: 22) to maximise the chance of the subjects knowing the cues. All JC2k cue words were selected from the second thousand JACET8000 word list.

ii. Potential cues were retained if the first association (reported by the online EAT) generated did not exceed 25% of responses as a potential cue "which typically generate a wide variety of different responses" (Meara and Fitzpatrick 2000: 22)

iii. Cues were retained if at least half of the most common EAT responses were not included in the first thousand JACET8000 word lists, to "give the testee a reasonable opportunity to generate a wide range of response words" (Meara and Fitzpatrick, 2000: 23)

The following then describes the selection process of the alternative (JC1k) cue words. First, an item was randomly selected from the first thousand (JACET8000) word list, Meara and Fitzpatrick's first selection criterion, by selecting every thirtieth item from the list. If the thirtieth item did not satisfy the criteria for selection, I selected the next

thirtieth item from the JACET8000 word list, and so on. In this example, the word *'brush'* was selected. In order to determine whether the potential cue '*brush'* could be included in the JC1k list I needed to go through the same rigorous selection procedure as Meara and Fitzpatrick (2000). Meara and Fitzpatrick selected their cues from Nation's word lists (1984) and used the same Edinburgh Associative Thesaurus (Kiss *et al.*, 1973). The cues were assessed using the online version of the Edinburgh Associative Thesaurus (Kiss *et al.* 1973) (http://www.eat.rl.ac.uk/) which produces lists of native speaker associations in response to potential cues. *'Brush'* was first entered into the online EAT to access a list of native speaker associations (table 5.1). Table 5.1 shows that the first response to the potential cue '*brush'* is *comb* and accounts for 17% of all native speaker associations. As the first response accounts for less than 25% of all responses, *brush* satisfies Meara and Fitzpatrick's (2000: 22) second selection criterion, and is accepted on the understanding that it might elicit a broad range of responses.

Table 5.1 Edinburgh Associative Thesaurus responses (Kiss *et al.*, 1973) for the stimulus word *'brush'*.

1. COMB 17%	15. TAR 2%	29. LONG 1%
2. HAIR 11%	16. WOOD 2%	30. MAT 1%
3. PAINT 8 %	17. ASIDE 1%	31. PAD 1%
4. TOOTH 4%	18. AUSTRALIA 1%	32. ROUGH 1%
5. BASIL 3%	19. BORDER 1%	33. SCRUBBING 1%
6. FOX 3%	20. CLEAN 1%	34. SHAFT 1%
7. OFF 3%	21. CLOTHES 1%	35. SHOE 1%
8. PAN 3%	22. DUST 1%	36. SOCKS 1%
9. SHOVEL 3%	23. ELECTRIC 1%	37. SQUIRREL 1%
10. SWEEP 3%	24. FLOOR 1%	38. TAKE 1%
11. UP 3%	25. FLUSH 1%	39. TEETH 1%
12. BROOM 2%	26. FOR 1%	40. THRUSH 1%
13. HANDLE 2%	27. HAIRS 1%	41. TREES 1%
14. TAIL 2%	28. HEAD 1%	42. WELL 1%
		43. WIPE 1%

I ignored all words that elicited no more than 3% of responses, such as *basil, fox, off, pan* and so on, and not including *tooth,* listed in table 5.1 above. The resultant list, shown in table 5.2 below, shows the most common responses made up of words produced by 4% or more of the EAT native speaker subjects' associations.

Table 5.2 Common Edinburgh Associative Thesaurus responses to *'brush'*.

1. COMB 17%
2. HAIR 11%
3. PAINT 8 %
4. TOOTH 4%

The frequencies of the common responses to the potential cue word were then examined, see table 5.2. If at least half of the most common responses were not included in the first thousand (of the JACET8000) word list, Meara and Fitzpatrick's third selection criterion, the word is included in the alternative list of 30 JC1k cues because it has the potential to

elicit infrequent items. From table 5.2 the word *paint* is the only word from the list above that occurs in the JACET8000 first thousand word list. Since the other words (*comb, hair* and *tooth*) are not included in the JACET8000 first thousand word list, '*brush*' meets all the criteria for inclusion and is therefore included in the JC1k set of cues because it is considered to have potential to elicit varied and infrequent responses. This process was repeated until 30 items for JC1k had been identified that matched the three criteria for selection. The procedure for the selection of the JC2k cues was essentially the same as the process for the JC1k list with the only difference being that the JC2k cue words were selected from the JACET8000 second thousand word list (while the JC1k cues were selected from the JACET8000 first thousand word list).

The procedure described above resulted in two new 30-item cue lists, JC1k and JC2k, which were presented to subjects along with the original Meara and Fitzpatrick (2000) cue list, Lexorig. Figure 5.1 below shows the three different sets of cues.

Figure 5.1 Three different sets of cues to test Lex30: Lexorig, JC1k, JC2k.

Lexorig	JC1k	JC2k
1. attack	1. away	1. affect
2. board	2. blow	2. area
3. close	3. brush	3. balance
4. cloth	4. chance	4. boundary
5. dig	5. common	5. cement
6. dirty	6. dance	6. comment
7. disease	7. district	7. connect
8. experience	8. ever	8. court
9. fruit	9. famous	9. degree
10. furniture	10. flag	10. dismiss
11. habit	11. get	11. energy
12. hold	12. head	12. extreme
13. hope	13. insect	13. flow
14. kick	14. knee	14. goal
15. map	15. list	15. hook
16. obey	16. mat	16. index
17. pot	17. mountain	17. just
18. potato	18. oil	18. load
19. real	19. pattern	19. memory
20. rest	20. policeman	20. oblige
21. rice	21. public	21. pain
22. science	22. religion	22. point
23. seat	*23.* secret	23. profession
24. spell	24. shirt	24. reaction
25. substance	*25.* sorry	25. research
26. stupid	26. smell	26. sale
27. television	27. spirit	27. ship
28. tooth	28. surprise	28. sport
29. trade	29. telephone	29. suit
30. window	30. tool	30. tight

5.2.2 Subjects.

The subjects were 40, L1 Japanese, undergraduate medical students enrolled in English as a foreign language (EFL) courses, aged eighteen, and made up of sixteen females and twenty-four males. The students took one and a half hours of English listening classes a week within the university, which took the form of listening to and discussing general academic English topics in a language laboratory. The subjects' TOEFL scores ranged from 440 to 530, and they were at a pre-intermediate level of speaking ability.

5.2.3 Method.

The subjects took Lexorig, JC1k, and JC2k at two different test times, six weeks apart. At each test time, the subjects were given the three Lex30 tests in turn (Lexorig, JC1k, and then JC2k). The three tests were as follows:

- Lexorig - Lex30 cue words (Meara and Fitzpatrick 2000) from Nation's (1984) first thousand word list
- JC1k - 30 cues from the JACET8000 first thousand list
- JC2k - 30 cues from the JACET8000 second thousand list

The tests were presented in written format (as in the original Meara and Fitzpatrick 2000 test). The subjects were given approximately 15 minutes to complete each Lex30 format. In order to avoid any potential overload, the subjects were given a five-minute break between each format at each test time.

Six weeks later the same group of subjects completed the same three tests under the same test conditions as the first test event. This six-week gap was given to minimize any practice effect and to allow for sufficient forgetting time that may otherwise have unduly skewed the scores. In between the two test times, the subjects had five 90-minute classes of English as part of their curriculum. All tests were then processed in the same way as Meara and Fitzpatrick (2000:24) (as in section 3.2.2) but using the JACET word lists to

score infrequent words. Fitzpatrick and Meara (2004: 71) claimed that the JACET8000 (Jacet 2003) word list might provide a more accurate, and up to date set of scores, and so the current test uses the JACET 8000 word lists (see section 4.2.2). All subject responses were typed into a machine-readable text file and scored by awarding one point for each infrequent (non 1k) item. Subject's scores were then compared. (Appendix 4 gives an example of a subjects' completed tests).

5.2.4 Results.

Table 5.3 shows the mean number of words produced in each task.

Table 5.3 Mean number of words produced in each task.

	Test time 1	Test time 2
Lexorig	110	115
JC1k	114	117
JC2k	115	110

Percentage scores, which "represent the number of infrequent words produced, as a percentage of the total number of responses given by that subject" (Fitzpatrick and Meara 2004: 56), were calculated in order to minimize the influence corpus size had on each subject's test score. Thus, all scores reported below are percentage scores. The results are presented in relation to the three research questions.

5.2.4.1 Will a different, but selected according to the same criteria, set of cue words produce similar results to the Lex30 original?

Table 5.4 shows the subjects' mean Lex30 scores and standard deviations from the original Lex30 test (Lexorig) and the JC1k test. In order to address the reliability question (see 5.3.2 below), subjects took the tests at two different times. I was therefore able to make two comparisons.

Table 5.4 Comparing means and standard deviations of Lexorig and JC1k scores.

Test time	Task	Mean	sd
Test time 1	Lexorig	24.1	8.6
	JC1k	23.3	7.9
Test time 2	Lexorig	28.3	9.1
	JC1k	26.5	7.6

As table 5.4 shows, the subjects mean scores are higher for the Lexorig test at both test times, although the differences in scores are relatively small (differences of 0.8 and 1.8 between means at respective test times). The slightly higher standard deviation for Lexorig shows that the subjects' scores are more differentiated than the JC1k test.

A comparison of means at the two test times, shown in table 5.5, indicates that there is no significant difference between Lexorig and JC1k mean scores at either test time.

Table 5.5 Paired t-scores for Lexorig and JC1k at each test time.

	Task	t-value	Sig
Test time 1	Lexorig – JC1k	0.785	0.437
Test time 2	Lexorig – JC1k	1.516	0.138

Table 5.6 below shows that there is a significant correlation between Lexorig and JC1k scores at both test times.

Table 5.6 Correlations between Lexorig and JC1k.

	Tasks	Correlation	Sig.
Test time 1	Lexorig and JC1k	0.720	<.001
Test time 2	Lexorig and JC1k	0.631	<.001

5.2.4.2 Will cue words from different frequency bands produce different scores?

The second research question asked whether responses to two sets of prompt words, from two different frequency bands (the first and second thousand frequency band of the JACET8000 word lists), would produce different scores. By comparing scores from tasks using the first and second thousand frequency bands, we can see if the scores are different and determine whether more frequent cues elicit a greater number of responses that are more frequent, and whether less frequent cues elicit a greater number of less frequent responses (Meara 1983: 29). Table 5.7 shows that the JC2k cues produced slightly lower Lex30 mean scores than the JC1k cues at test time one and at test time two. The standard deviations are lower for the JC2k cues, suggesting that subjects' scores are more clustered when presented with the lower frequency (JC2k) cue words.

Table 5.7 Comparing means and standard deviations of JC1k and JC2k scores.

Test time	Task	Mean	Sd
Test time 1	JC1k	23.3	7.9
	JC2k	21.7	5.5
Test time 2	JC1k	26.5	7.6
	JC2k	24.3	5.9

The t-values in table 5.8 show the difference between the JC1k and JC2k tasks is small but the difference is significant at test time two ($p<.05$), tending towards significance at test time one. The t-values indicate that something slightly different happens when subjects are faced with cues from the second thousand frequency band. One possible explanation is that the subjects might not have been familiar with some of the words used as cues from the second thousand band and therefore produce fewer responses. Alternatively, subjects might have produced different kinds of responses to the 2k cues, such as more frequent responses.

Table 5.8 Paired t-scores for JC1k and JC2k at each test time.

	Test	t-value	Sig
Test time 1	JC1k- JC2k	1.913	0.06
Test time 2	JC1k- JC2k	2.417	0.02

The correlations between JC1k and JC2k shown in table 5.9 are significant and indicate that the two tests are working in broadly the same way. The significant correlations appear to suggest that the JC2k cues affect all of the subjects in a similar way.

Table 5.9 Correlations between JC1k and JC2k.

	Tasks	N	Correlation	Sig.
Test time 1	JC1k - JC2k	40	.739	<.001
Test time 2	JC1k - JC2k	40	.665	<.001

The reason for the second research question was to determine whether cues selected from the second thousand frequency band would produce systematically different scores and the results appear to show that they do not.

5.2.4.3 How, if at all, do individual subjects' Lex30 (with Lexorig, JC1k, JC2k cues) scores change over a 6-week period?

This section reports the results of the test-retests of the three versions of Lex30 at test time one and test time two. Table 5.10 shows the means of the three different Lex30 test scores. The means at test time 2 are consistently higher than the tests taken at test time 1.

Table 5.10 Mean scores and standard deviations of scores at test time 1 and test time 2.

	Test time 1 (sd)	Test time 2 (sd)
Lexorig	24.1 (8.6)	28.3 (9.1)
JC1k	23.3 (7.9)	26.5 (7.6)
JC2k	21.7 (5.5)	24.3 (5.9)

The t-values shown in table 5.11 show the significant difference in means between subjects' scores at test time one and test time two. Two possible reasons for this are, first, that the difference in scores may be due to the subject's vocabulary growth (because of the intervening five classes of English between the two test times) and second, because of the practice effect.

Table 5.11 Paired t-scores for each test pair.

Test times 1 and 2	t-value	Sig
Lexorig	4.923	<.001
JC1k	6.967	<.001
JC2k	7.111	<.001

Table 5.12 shows that each mean test score generated strong and significant correlations for the two test times. These strong and significant correlations indicate good test-retest reliability.

Table 5.12 Correlations between scores on the same test versions at different test times.

Test times 1 and 2	Correlation	Sig
Lexorig	0.814	<.001
JC1k	0.929	<.001
JC2k	0.916	<.001

The correlations between test and retest scores are shown in figures 5.2-5.4. The figures show clearly that scores improve from test time one to test time two for each of the three Lex30 formats.

Figure 5.2 Test time two compared with Test time one: Lexorig.

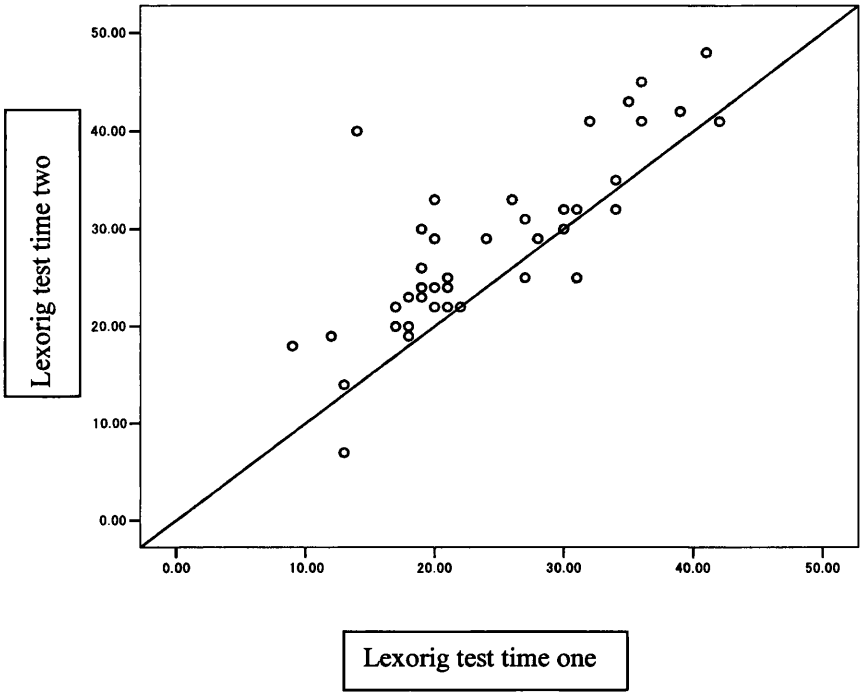


Figure 5.3 Test time two compared with Test time one: JC1k.

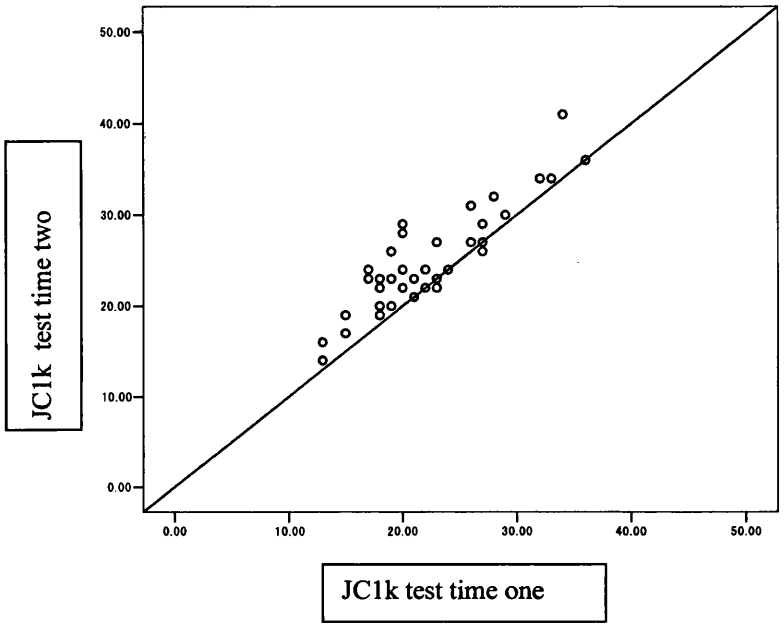
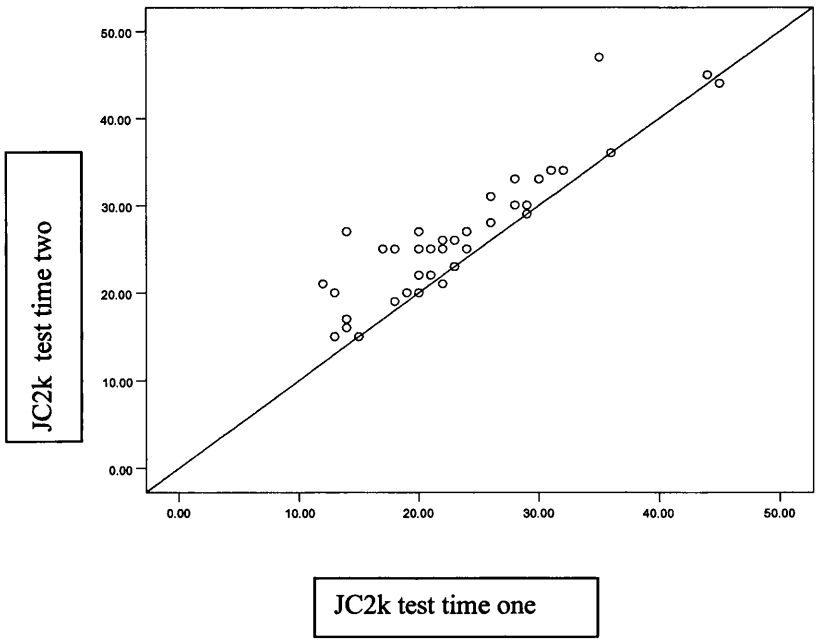


Figure 5.4 Test time two compared with Test time one: JC2k.



5.3 Discussion.

The experiment raises three factors to consider, which I address in the following discussion.

The first factor relates to the first research question which asked whether different cues, but selected according to the same selection criteria, would produce similar results to the Lex30 original. I compared Lex30 scores from two different sets of cue words from the first 1000 frequency band (Lexorig and JC1k). Table 5.3 shows that the difference between the two mean test scores is relatively small (a difference of 0.8 at test time one, and a difference of 1.8 at test time two). The strong and significant correlation (0.720 $p<.001$) between Lexorig and JC1k appears to suggest that the two are broadly predictive of one another. The strong and significant correlations indicate that with different cues from the same (1k) frequency band the two tests elicit broadly similar results, suggesting that the two sets of cues are tapping the same kind of knowledge.

The second factor relates to the second research question, which asked whether cue words from different frequency bands, (the first, and second thousand frequency bands) elicit similar scores. The t-scores shown in table 5.7 suggest a small difference between JC1k and JC2k (test time one: JC1k-JC2k $t=1.913$, $p=.06$; and, test time two: JC1k-JC2k $t=2.417$, $p=.02$). The smaller standard deviations with the JC2k cues appear to support continued use of the Lexorig cues (selected from the first thousand-word list) because the standard deviations for the JC2k test suggest a tighter grouping of subject performance on the task compared to the standard deviations for both Lexorig and JC1k. While there is not that much difference between the standard deviations, they tend to suggest that the cues taken from the first thousand word list, Lexorig and JC1k, show greater variation between different subject proficiencies, compared to the Lex30 task with cues taken from the second thousand word list. The correlations shown in table 5.8 suggest that the two tests are working in broadly the same way and that the lower frequency words did not necessarily provoke lower frequency responses.

The third factor relates to the third research question which examined the potential reliability of Lex30 as a test and asked whether individual subjects' scores change over the test period. The correlations (table 5.12) show strong and significant patterns of correlation between test times for each of the three tasks. The mean scores (table 5.10) indicate that subjects' scores are consistently higher at test time two and the t-values (table 5.11) show a significant pattern of difference between test time one and test time two. These analyses suggest that the Lex30 test-retest scores indicate an improvement in subjects' abilities to produce vocabulary in response to the Lex30 cues over the six-week period. The six week between test-retest period was chosen in order to minimize any 'practice effect' (Bachman, 1990: 182) and as the subjects had five 90-minute classes in between the two test times, the classes included a significant amount of vocabulary instruction, the improvement appears more likely to be due to the teaching intervention rather than the test effect. This section tentatively concludes that Lex30 is consistent as a test measurement.

The introduction to this chapter suggested that if the subjects achieved similar scores from two different sets of Lex30 cues, selected according to the same criteria as Meara and Fitzpatrick, then this might contribute to the construct validity of the test. The t-tests indicate that there was no significant difference between the two tasks (Lexorig-JC1k $t=0.785$ $p= 0.437$). The strong correlations between the Lexorig scores and JC1k scores indicate that the two tests (JC1k and Lexorig) are to some extent predictive of each other and suggests that Lex30 appears to have construct validity.

5.4 Conclusion.

The results from chapter five suggest that regardless of the frequency band from which the cues were selected, Lex30 appears successful at eliciting responses that appear, broadly, to be reflective of subjects' ability to produce infrequent vocabulary items. The cues were selected from two different frequency bands for different versions of the task

(the first and second thousand frequency bands). Accordingly the study tested with three different versions of Lex30 in order to elicit subjects' productive vocabulary knowledge with three different sets of cues: Lexorig (Meara and Fitzpatrick's (2000) original cues), JC1k (30 cues selected from the first thousand frequency band), and JC2k (30 cues selected from the second thousand frequency band). The mean scores from each of the three sets of cues (Lexorig, JC1k, and JC2k) were similar, suggests that regardless of the cues we test with, Lex30 appears to elicit similar scores that elicit similar proportions of infrequent vocabulary items. The results from chapter five also suggest that the subjects' ability to produce infrequent items improves over the extended (six-week) test period. This improvement tentatively indicates validity and, one might extrapolate, the increased proficiency is potentially due to the teaching intervention rather than the test effect.

Chapter five also found that different Lex30 cues elicit similar results that appear to indicate that subjects produce similar proportions of infrequent items in response to different Lex30 cues. The results imply that the test has some internal validity and cautiously indicates that it might now be time to explore issues raised by Fitzpatrick (2007) regarding performance on Lex30 against other tests that claim to measure productive vocabulary.

Chapter 6 Comparing two measures of free productive vocabulary: Lex30 and the Lexical Frequency Profile

6.1 Introduction.

The comparisons in chapter five are based on parallel forms of the Lex30 test taken at two test times. The comparability between the different test formats, and the increased scores after a study period, indicate a degree of construct validity. In this chapter, I turn to address concerns raised by Fitzpatrick (2007) (section 2.3.3) who suggests that we need to compare Lex30 with other measures in order to determine whether it elicits productive vocabulary knowledge. As Bachman suggests, a comparison of "[an] individuals' performance on another test of the [same] ability in question" (1990b: 248) would support the concurrent validity of a test if the test scores correlate. In this chapter, because both Meara and Fitzpatrick (2000: 19) and Laufer and Nation (1999: 37) claim their tasks measure 'productive' or 'free active' vocabulary, I compare scores from Lex30 with scores from the LFP to test the concurrent validity of Lex30.

Fitzpatrick (2007) attempted to address Baba's (2002) belief that "it would be [also] profitable to compare Lex30 with other tests designed to assess a similar construct" (2002: 69) by comparing Lex30 with two other tests: The Productive Levels Test and a translation task. Fitzpatrick selected these two tests because they "share certain characteristics with Lex30" (Fitzpatrick 2007: 122). The three tests (Lex30, the Productive Levels Test, and the translation task) have the following five characteristics in common, they all: i) operate on the assumption that vocabulary can be measured; ii) claim to measure productive vocabulary knowledge; iii) are paper tests and take a short time to complete; iv) use frequency bands; and, v) measure vocabulary knowledge, not syntactical knowledge.

Fitzpatrick found relatively weak but significant correlations between Lex30 and the Productive Levels Test (0.504 ($p < .01$)) and with the translation task (0.651 ($p < .01$)). Meanwhile, the correlation between the Productive Levels Test and the translation task

was much stronger (0.843 ($p < 0.1$)) because, Fitzpatrick suggests, the Productive Levels Test and the translation task might be measuring different aspects of knowledge to those measured by Lex30. In other words, although the test constructs of the Productive Levels Test and the translation task seem to have considerable overlap with those of Lex30, the results produced in practice by the Productive Levels Test and the translation task do not correlate nearly so strongly with those of Lex30 as with each other. For Fitzpatrick, this stronger correlation between the Productive Levels Test and the translation task is likely for three main reasons. First, both tasks are measuring word knowledge up to the 3000 word level (because subjects did not answer many questions from the 5k, UWL, and 10k parts of the Productive Levels Test, and the translation task only tested with items up to the 3000 word level). Second, both tasks require subjects to respond with a set of pre-determined words (whereas Lex30 gives credit for any infrequent item produced). Third, subjects encounter increasingly infrequent test items as they advance through the tests, and therefore the tests increase in difficulty (whereas the task difficulty is constant throughout the Lex30 task, in which subjects provide up to four items for each of the 30 cues). Such differences lead Fitzpatrick to question whether all three of the tests are accessing different abilities. She argued that we ought to be examining whether tests have different 'activation properties' (2007: 127), including, for instance, whether a test activates knowledge of how a word is written or spelled, or activates knowledge of the patterns in which we ought to use the word (Nation 1990: 27).

As the LFP heralds itself as a test of productive (or 'free active') vocabulary (Laufer and Nation 1999: 37) the same could also be claimed for Lex30. Laufer and Nation state "we refer to the ability to use a word at one's free will as free productive ability. This type of knowledge is measured by the Lexical Frequency Profile" (Laufer and Nation 1999: 37). The treatment of language produced in response to the LFP is comparable to the treatment of language produced in response to Lex30: misspellings are corrected, incorrect derivatives are considered acceptable as derivatives from one family of the same frequency band, and all proper nouns are deleted. Based on such similarities between the two tasks, this chapter compares Lex30 scores with scores from 'another

test of the ability in question' (Bachman 1990b: 248): the Lexical Frequency Profile (LFP). One difference between the two tasks, though, is the definition of infrequent items. Although both the LFP and Lex30 are interested in the proportion of infrequent words produced, Lex30 defines infrequent as outside the first thousand frequency band, Laufer and Nation (1995) suggest scores can be obtained by taking infrequent words as being outside the second thousand frequency band. This chapter reports on two comparative studies, with the findings from study one informing study two. I first report on my study comparing Lex30 with the LFP (reviewed in detail in section 2.2.6). The second study, based on the findings of the first, reports on a study in which the LFP is adapted to better match the construct of Lex30.

6.2.1 Study one - Comparing Lex30 and the LFP.

I have chosen to compare Lex30 with the LFP on the basis that their respective authors (Meara and Fitzpatrick 2000; Laufer and Nation 1995) each claim that their test measures productive vocabulary. Meara and Fitzpatrick describe Lex30 as a "test of productive vocabulary" (2000:19), and Laufer describes the LFP as a "tool which attempts to measure free productive vocabulary" (2005: 582). Both tests use frequency list calculations and both score approximately spelled items. Accordingly, the aim of the first study is to examine whether scores correlate on the different tests in order to determine whether the test scores are largely predictive of each other. The research question for study one is:

Is the proportion of infrequent items derived from Lex30 different from that derived from a discursive task?

6.2.1.1 Subjects.

The test subjects were 80 L1 Japanese university students. The subjects came from two different faculties at Osaka University (medical and engineering science). The students were aged between eighteen and twenty-one, and made up of twenty-six females and

fifty-four males. The subjects had three hours of English speaking classes each week, the classes took the form of students discussing general social issues. Scores from an independent TOEFL test (ranging from 420 to 480) indicated that the subjects' proficiency levels were clustered around the pre-intermediate to intermediate proficiency levels.

6.2.1.2 Method.

The tests took place over a two-week period. In the first week, subjects took the written format of Lex30 and the first of the two LFP composition tasks. In order to avoid overload, subjects had a five-minute break between Lex30 and the first LFP task. In between the two class periods, the subjects had no English classes. In the following class session, the subjects completed the second LFP composition task.

6.2.1.3 Lex30.

The criteria for scoring each set of subject responses were similar to those of Meara and Fitzpatrick (2000) but with the difference being that the scores were instead processed online using Cobb's WebVP (http://www.lextutor.ca/vp/eng/). I used the WebVP because it uses the same word lists according to which Laufer and Nation (1995) scored the Lexical Frequency Profile. The WebVP contains words from: the first thousand most frequent word families; the second thousand word list; the academic word list (AWL); and, a list of words that do not appear on the other lists (Off-list words). Using the online scorer was a time efficient way to score because it was relatively straightforward to identify the items produced within their respective frequency bands. Responses were processed before scoring according to the established Lex30 protocol (see sections 3.2.2 and 3.2.3 for examples). Each subject's paper test was typed into a computer text file and then categorized using the WebVP. The WebVP enables users to paste text into an online window, click 'submit window,' and the text is then categorized according to the number of items produced within each of the four frequency levels (1k, 2k, AWL, Off-list words). The WebVP categorizes proper nouns as 1k words, and separates types and

tokens. I scored only using the total number of types, as this counts all the different words, thus avoiding counting a word produced repeatedly. All proper nouns and grammar words were scored as 1k, as in Meara and Fitzpatrick (2000). Each subject's Lex30% score was the total number of infrequent responses, consisting of all items produced outside the first 1000 band (i.e. Lex30 score= 2k+AWL+Off-List words), as a proportion of the total number of items produced.

6.2.1.4 The Lexical Frequency Profile.

The following describes the procedure taken in administering the LFP task. In the same way as in Laufer and Nation (1995), the subjects wrote two compositions, one per week in two class periods, a week apart. Each composition task was completed within one hour, and the length of each composition was limited to 300 -350 words. These criteria were based on Laufer and Nation's (1995) contention that profiles of less than 200 words were found to be unstable and that writing 300-350 words was not unfeasible in the one-hour permitted. The compulsory first question from Laufer and Nation (1995) is:

- *'Should a government be allowed to limit the number of children a family can have?' Discuss this idea considering basic human rights and the danger of population explosion.*

For the second question (Laufer and Nation 1995), subjects had to choose one of the following three questions to which to respond.

- *'A person cannot be poor and happy, because money is always needed to gain something that is important to that person'. Argue for and against this idea;*
- *'It is always what you do not have as a child that is important to you as an adult'. Agree or disagree with this statement;*
- *'In a free country, industry has the right to develop any product that will sell, and industry can sell it to anyone who can pay for it'. Do you agree with this*

idea or do you think that the government should be able to control what is produced and sold?

To ensure motivation, in the same way as Laufer and Nation (1995), the subjects were told that the scores for their compositions were to be included in their final course grades. The subjects' compositions were processed using the same four criteria as Laufer and Nation (1995): i) if a word was clearly used incorrectly, it was omitted and not considered part of a subject's productive lexicon; ii) misspellings were corrected; iii) incorrect derivatives were considered acceptable as derivatives from one word family of the same frequency; and, iv) all proper nouns were deleted from compositions.

As with Lex30, the LFP texts were typed into a computer text file. Each text was then copied and pasted into the WebVP (http://www.lextutor.ca/vp/eng/), to be categorized according to the number of items produced within each of the four frequency levels (1k, 2k, AWL, Off-List). The WebVP uses the same word lists as in Laufer and Nation's (1995) original paper. In order to make comparisons between the LFP and Lex30, an LFP mean score was calculated. Given the obvious statistical difficulties in comparing a Lex30 'score' with an LFP 'profile' an LFP score was instead calculated based on the proportion of responses within the AWL and Off-List bands (i.e., the total proportion of words produced in the AWL+ Off-List frequency bands= LFP score), following Laufer and Nation (1995:312). As an example, table 6.1 shows that the LFP score for subject 1 was 8.69% (the total proportion of words produced in the AWL+Off-List bands (7.02%+1.67%=8.69%)). A mean LFP score was then calculated in order to compare the single Lex30 score with the two LFP scores (one for each composition (as Laufer and Nation (1995: 312)). The LFP mean score was taken as the average percentage of responses subjects produced within the AWL and Off-List frequency bands, for the two composition tasks.

Table 6.1 Lexical Frequency Profile score for subject 1.

	1000	2000	AWL	Off-List	Score
Subject 1	89.64%	1.67%	7.02%	1.67%	8.69%

6.2.1.5 Results.

The research question asked whether the proportion of infrequent items derived from Lex30 was different from that derived from a discursive task, and so the following section compares the Lex30 and LFP scores. The scores indicate that the subjects responded to the two tasks very differently in terms of the infrequent words that they produced. Table 6.2 shows the results of the Lex30 task. Given that subjects vary in terms of how many words they produce, and to minimise the effect of that variable, I use percentage scores for the rest of this analysis. Percentage scoring also makes sense because the LFP uses percentages not raw scores. The LFP scores are also given as percentages, as in Laufer and Nation (1995) (see table 6.3). Percentage scores allow a comparison of data independent of the total number of words produced when an objective comparison is required between, for example, a subject who has given 120 Lex30 responses and one who has given 90. The mean number of items produced for Lex30 was 115 per subject.

Table 6.2 Lex30 mean scores.

	Mean score *(sd)*	Min score	Max score
Lex30 %	43.63% *(5.89)*	29.41%	57.83%

Table 6.3 shows the percentage scores for the LFP for the two composition tasks. The correlation between the scores generated from the two LFP (LFP1 and LFP2) composition tasks, though low, was significant, and show that the LFP task generates broadly similar scores ($r=0.346$ ($p<0.01$)).

Table 6.3 Lexical Frequency Profile mean scores.

Mean	LFP% score (AWL+Off-list words) (sd)	Min LFP% score	Max LFP% score
LFP 1st %	5.2% (2.6)	0.70%	14.96%
LFP2nd %	5.37% (2.9)	0.37%	14.14%
LFP %	5.28% (2.3)	0.69%	11.22%

However, the correlation in table 6.4 indicates that subjects responded in very different ways to the LFP and Lex30 tasks.

Table 6.4 Correlation between Lex30 percentage and the Lexical Frequency Profile percentage scores.

	Lex30%
LFP% mean	0.186 p=.098

The correlation shown in table 6.4 illustrates that the Lex30 percentage scores are not predictive of the LFP% scores. The scores show that subjects produced a greater proportion of infrequent items in response to the Lex30 task (mean Lex30 percentage score: 43.63%) than the LFP task (mean LFP percentage score 5.28%). The lower LFP scores might be somewhat attributable to the elicitation process and the scoring, but the fact that the scores do not correlate does indicate that the two tasks are measuring productive vocabulary ability differently or at least with different degrees of accuracy. There are three possible explanations for the lack of correlation and different scores. First, the LFP discursive task necessarily produces many function words, which influences the proportion of infrequent items subjects appear able to produce, while Lex30 produces very few or almost no function words. Second, a subject's ability to

produce infrequent vocabulary items might be limited by the discursive aims of the LFP. Third, and potentially most importantly, the Lex30 awards points for items outside 1k, while the LFP task only awards points for everything outside 2k. In order to address these three issues, the Brainstorm Frequency Profile was devised in study two and is explained below.

6.2.2 Study two - Comparing Lex30, the Brainstorm Frequency Profile Task and the LFP.

The lack of significant correlation between Lex30 and the LFP in the first study was perhaps surprising, given that the two tests both seek to measure 'free active' (Laufer and Nation 1999) or productive vocabulary. One of the reservations about the LFP (as discussed in detail in 2.2.6) was that the focus on the composition, and the composition skills required, as well as the use of function words might limit subjects opportunities to provide a large proportion of infrequent items. The results from study one indicate that these are problematic issues and pertinent concerns especially when comparing LFP scores with seemingly superior Lex30 scores. I aim to address these issues in this second study.

The results from study one seem to support Fitzpatrick's suggestion that the different performances on the different tests may be due to their 'different activation properties' (2007: 127). Thus, the contrasting ways in which the LFP and Lex30 aim to elicit productive vocabulary knowledge may be worth examining in detail. As discussed in the reviews of Laufer and Nation (1995) (section 2.2.6), and Meara (2005) and Laufer (2005) (section 2.2.7), the LFP task appears to make it less likely that subjects will be able to produce so many infrequent items, and this is probably due to the demands of the composition task.

Thus, my aim in this second study is to adapt the LFP in a way that eliminates or at least greatly attenuates the influence of the discursive nature of the task. I wanted to retain the element of the subjects having to respond to a composition question, but to avoid them

having to produce full sentences in doing so. The second study in this chapter therefore introduces the Brainstorm Frequency Profile. Comparisons between Brainstorm Frequency Profile scores and LFP scores might reveal the extent to which subjects are limited by the standard LFP composition task. The rationale in designing the Brainstorm Frequency Profile task is to determine whether a different score will be achieved if vocabulary is elicited in the form of a brainstormed word list as opposed to the discourse generated by the LFP composition task. Although the Brainstorm Frequency Profile task does not require subjects to respond with a discursive essay it does nevertheless use the same question prompt as the LFP. The Brainstorm Frequency Profile works by asking subjects to brainstorm their responses to Laufer and Nation's (1995) first compulsory LFP question (*'Should a government be allowed to limit the number of children a family can have?' Discuss this idea considering basic human rights and the danger of population explosion*). This first LFP compulsory question was chosen for the Brainstorm Frequency Profile task because the question seemed sufficiently general and because there was little difference between the scores from the two LFP composition questions from the first study (see table 6.3). The significant correlation between the scores generated from the two LFP (LFP1 and LFP2) composition tasks, shows that the LFP task generates broadly similar scores. For the ensuing experiment in this second study, the subjects responded to the Brainstorm Frequency Profile task by brainstorming their responses to the first LFP question, and then writing compositions in response to the second LFP task in the standard LFP composition format. Thus, the second study compares scores from three tasks: Lex30, the Brainstorm Frequency Profile task, and the second LFP composition task. Accordingly, the research question for study two is:

Is there a closer relationship between Brainstorm Frequency Profile and Lex30 scores than between LFP and Lex30 scores?

6.2.2.1 Subjects.

The subject group for the second study were selected based on the perceived similarities with the subject group from the first study. As in the first study, the participants were 80 Japanese L1 university students. These subjects were from two faculties at Osaka University: engineering and technology. The students were aged between eighteen and twenty, and made up of eight females and seventy-two males. The subjects had three hours of English speaking classes each week, the classes took the form of students discussing general social issues. Scores from an independent TOEFL test (ranging from 410 to 470) indicated that the subjects' proficiency levels again ranged from pre-intermediate to intermediate (the TOEFL scores for the subjects from study one ranged from 420 to 480, so their proficiency levels were almost identically clustered around pre-intermediate to intermediate levels).

6.2.2.2 Method.

The subjects undertook three tasks: Lex30, the Brainstorm Frequency Profile task, and the second LFP composition question. In the first week of testing, the subjects took Lex30 and the Brainstorm Frequency Profile task. The procedure for conducting the Lex30 task was the same as the standard Lex30 protocol (described in full in section 3.2.2). Subjects had a five-minute break between the Lex30 task and the Brainstorm Frequency Profile task in order to avoid potential overload and to avoid performance in one task influencing the other. In between the two class test periods, the subjects had no classes of English. In the second week of testing, the subjects completed the LFP composition task in the one hour permitted (Laufer and Nation 1995), and it was then processed as in study one.

6.2.2.3 The Brainstorm Frequency Profile task.

The following describes the procedure taken for the Brainstorm Frequency Profile task in which the subjects brainstormed their responses to the first LFP question in the form

of single words. The subjects were asked to give as many one-word responses as they could to the Brainstorm Frequency Profile question. An example was given before starting, with a different question, in order to highlight that subjects were not required to write their responses in the form of a composition. Figure 6.1 shows the instructions and an example response (from subject 1) for the Brainstorm Frequency Profile task. The original LFP question cue was changed from 'Discuss' to the Brainstorm Frequency Profile task question cue of 'Write as many one-word responses as possible to'. I changed the question cue in order to make it very clear to the subjects that they were to brainstorm their responses rather than respond in composition form (as is required by the standard LFP task).

Figure 6.1 Instructions and example response for the Brainstorm Frequency Profile.

'Should a government be allowed to limit the number of children a family can have?' Write as many one-word responses as possible to this idea considering basic human rights and the danger of population explosion.

government limit number children family father mother
brother sister cost income study elementary junior high
school college university job club lesson agriculture cultivate
soil village vegetable carrot eggplant tomato potato money
economy breakfast lunch dinner population rights birthday
friend party cake food clothing shoes hair hobby baseball
tennis soccer wash drink glove bat ball racket noisy

In order to replicate Laufer and Nation's (1995: 312) test conditions as closely as possible, I told the subjects that their LFP (and Brainstorm Frequency Profile) task scores would be included in their final course grades. The Brainstorm Frequency Profile task scores were processed using the WebVP, and by counting all the words produced in the AWL and Off-list bands, the same as the LFP task. In the same way as the treatment

of responses to Lex30, all misspellings were corrected, items were lemmatized, and proper nouns were treated as 1k items.

6.2.2.4 Results.

The research question for study two asked whether there is a closer relationship between Brainstorm Frequency Profile task scores and Lex30 scores than between LFP scores and Lex30 scores. Table 6.5 shows the results of the Lex30 task. The Lex30% scores generated in this second study (mean 38.51% (sd 5.9)) are lower than the ones generated by the subjects from the first study (mean 43.63% (sd 5.9)) somewhat unexpectedly as the participants had similar TOEFL scores, although the subjects' TOEFL scores for study one were slightly higher. Both sets of scores, for study one and study two are notably higher than those reported in earlier chapters, and I explore reasons for this in the discussion section (6.3) that follows. The mean percentage scores generated by the Brainstorm Frequency Profile task (12.06% (sd 9.1)) are lower than those generated by Lex30 (38.11% (sd 5.85)), but higher than the LFP scores (mean 5% (sd 3.4)). The Brainstorm Frequency Profile task elicited higher mean scores than the LFP task.

Table 6.5 Mean scores (study two).

	Mean score (sd)	Min score	Max score
Lex30%	38.51% (5.9)	23.6%	47.6%
BFP%	12.06% (9.1)	0.0%	39.7%
LFP %	5% (3.4)	0.0%	15.5%

The far higher Lex30% scores than the Brainstorm Frequency Profile% task scores need explaining. The obvious explanation for the lower Brainstorm Frequency Profile% task scores is that the scoring is different. Lex30% scores include 2k words, but the Brainstorm Frequency Profile and LFP tasks do not because I was following Laufer and Nation's (1995) scoring system. Lex30% scores all words produced outside the 1k band

(i.e., 2k+AWL+Off-list bands), and the Brainstorm Frequency Profile% and LFP% score all words produced outside the 2k band (i.e., AWL+Off-list bands).

Table 6.6 Lex30%, Brainstorm Frequency Profile% and Lexical Frequency Profile% correlations.

	LFP%	BFP %
Lex30 %	0.167 p=.139	0.153 p=.175
BFP %	0.004 p=.970	

An examination of the correlations (shown in table 6.6) shows that the percentage scores do not correlate, intimating that the number of words produced might be an important factor. Lex30 and the BFP elicit a smaller maximum number of words (120) than the standard LFP task (300), and the BFP might have elicited even fewer items in total than Lex30. The smaller number of items produced might have influenced the percentage scores and might explain the lack of correlation between the three tasks. The lack of strong and significant correlations between the different task scores might be due to the different frequency bands used to score the different tasks.

This different way of counting infrequent items was one potential reason for the lack of a significant correlation between Lex30 and the LFP discussed in study one. In order to make better comparisons with Lex30, I rescored the Brainstorm Frequency Profile task and LFP data using the same frequency bands used to score the Lex30 task (2k+AWL+Off-list bands), so that all three tests defined infrequent items in the same way. The following is a reanalysis of the same raw data but with the scoring of infrequent items taken as all those produced within the 2k, AWL, and Off-list bands, for the three tasks.

Table 6.7 Lex30, BFP and LFP mean original scores and 2k+AWL+Off-list scores.

	Mean original score (sd)	Mean (2k+AWL+Off-list) score (sd)
Lex30	38.51% (5.9)	38.5% (5.9)
BFP	12.06% (9.1)	28.1% (9.1)
LFP	5% (3.4)	9.1% (4.7)

Table 6.7 shows this reanalysis of the scoring to reflect all three tests now defining infrequent items in the same way as Lex30. The Brainstorm Frequency Profile task and LFP mean scores are predictably higher than how they had been originally scored. Both the LFP and BFP scores have increased, especially the BFP mean score, having more than doubled. The reanalysis of the scoring shows that, when we include items produced within the 2k frequency band, Brainstorm Frequency Profile% mean scores are notably higher than the scores tallied using only the items produced in the AWL+Off-list bands. In addition, the inclusion of the 2k frequency band shows that subjects produce a far greater proportion of infrequent items in response to the Brainstorm Frequency Profile task than the LFP task. Lex30% mean scores (38.1% (sd 5.8)) are still higher than the mean percentage scores generated by the Brainstorm Frequency Profile% task (28.1% (sd 9.1)). This reanalysis of the scores, having redefined 'infrequent' items, appears to indicate that the choice of task can influence the eliciting of infrequent items.

Table 6.8 Lex30, Brainstorm Frequency Profile and Lexical Frequency Profile (scored using 2k+AWL+Off-list) correlations.

	LFP%	BFP %
Lex30 %	0.159 p=.159	0.056 p=.622
BFP %	0.181 p=.108	

Table 6.8 shows the correlations between the three tasks when all tests are scored using 2k+AWL+Off-list frequency bands. There is a lack of significant correlation between all of the tasks. The lack of a significant correlation between the Lex30 percentage scores and Brainstorm Frequency Profile task percentage scores is surprising (because both elicit lists of words and the scoring systems are similar), but suggests that subjects respond in different ways to the two tasks, a possibility that needs exploring. There is also a lack of significant correlation between Lex30% and Brainstorm Frequency Profile% scores when Brainstorm Frequency Profile% scores ar*et al*lied according to AWL+Off-list, shown above in table 6.6. Both Lex30 and the Brainstorm Frequency Profile appear to elicit infrequent items, but the contrasting cue types (30 Lex30 cues compared to the one BFP question cue) might explain the lack of significant correlation between the mean Brainstorm Frequency Profile% and mean Lex30% scores. Meanwhile, the LFP task in this second study did not correlate significantly with either the Lex30 or Brainstorm Frequency Profile tasks.

6.3 Discussion.

The research question for study two was devised to determine whether there was a closer relationship between Brainstorm Frequency Profile task and Lex30 scores than between LFP and Lex30 scores. The results from the first and second studies showed that none of the test scores correlate with each other, but the Brainstorm Frequency Profile elicited a greater proportion of infrequent items than the LFP and appears to have a closer relationship to Lex30 than the LFP. There are four issues that might have contributed to these findings, and these are explored below.

The first issue is to attempt to examine the difference between discursive and one-word responses. Tables 6.6 and 6.8 showed no significant correlation between the LFP and the other tasks. This section examines the lack of significant correlation by exploring the differences between the language provided for the LFP and that for the two other tasks. The results from the first study confirmed that the discursive LFP task elicited a much

smaller proportion of infrequent responses (LFP mean score 5.3%) than the Lex30 task (Lex30 mean score 43.6%). The results from the second study also show that the LFP task elicited a smaller proportion of infrequent responses (LFP mean score 5%) than Lex30 (Lex30% mean score 38.5%) and the Brainstorm Frequency Profile (BFP% mean score 12.1%). Even when scoring includes the 2k band, the LFP mean scores (9.1%) are still lower than the Brainstorm Frequency Profile task mean scores (28.1%). These lower LFP task scores show that the task is less successful at eliciting scoring items from subjects compared to the two other tasks, and these results emphasise that the choice of task influences the eliciting of infrequent items. The lower LFP scores are likely because, as Meara (2005:34) observes (as discussed in 2.2.7), the LFP elicits data through the medium of a composition task.

The lack of correlation between test scores might be due to two particular features of the LFP task. First, subjects inevitably provide many function words in their LFP compositions. The results shown in table 6.7 indicate that, when responding to the Brainstorm Frequency Profile task (in which subjects brainstormed responses to the first LFP question) subjects demonstrate 'fuller' (see section 2.2.6) vocabulary knowledge than they are able to in the LFP task. When asked to respond to the brainstorm task as well as to the LFP composition task the subjects' scores indicate that it is the Brainstorm Frequency Profile task, that elicits the greater proportion of infrequent responses, thus enabling them to achieve higher scores. This outcome is probably because brainstorming does not ordinarily elicit function words.

Second, subjects might provide a greater number of more frequent content words on the LFP, knowing that they can fit them into wider contexts, rather than rarer, or less frequent (and therefore higher scoring) content words, which they cannot. This suggestion relates to the concern that the knowledge of syntax required by a comprehension task (Read 2000: 18) influences subject responses. Laufer and Nation claim that the LFP "has value as an indicator of quality of vocabulary use in that it can show the extent to which subjects are making the fullest use of their available

vocabulary knowledge" (1995:308). However, subjects might be concentrating primarily on syntax and may not, therefore, be making the 'fullest' use of their available vocabulary knowledge in the way that Laufer and Nation appear to claim. Subjects might also know a greater number of rare words but then avoid using them in response to the LFP task. The superior Brainstorm Frequency Profile task scores indicate that this reluctance to deploy the same proportion of infrequent items on the LFP might be the case. Subjects who take risks by providing a greater number of rarer words score more on the LFP than those who do not. Subjects can gain points for providing a greater number of rare words only if those words are presented as part of a text that adheres to the conventions of the composition task and fits within the broader context. For instance, those subjects who do well to provide the infrequent item 'vicissitudes' would only be provided the point if this word fits within a comprehensible and appropriate context. This scoring system suggests that not only do subjects need to know how to place a word in a broader context, demanding a working knowledge of the surrounding items, but they also need to demonstrate that they know the word and its applications quite well in order to produce it in response to the LFP task. They might also need to know the word's collocations and inflections, and so on. Hence, when subjects respond to the LFP task, they are probably far more likely to produce words that they know well, and that fit within the structure of the composition, than take the risk of providing more rare content words they may not know so well. In this sense, while subjects gain points for providing any infrequent item, they appear less likely to do so in response to the LFP task compared to the Brainstorm Frequency Profile task, probably because of the deeper knowledge that is needed to integrate a word accurately into discourse. Subjects need to know much more about a word to produce it in the LFP compared to the Brainstorm Frequency Profile task and Lex30, and the scores from the second study show that subjects produced a greater number of infrequent items in response to Lex30. Thus, Lex30 appears far more likely to elicit infrequent vocabulary items than the LFP.

This first issue of this discussion examined differences between the three tasks in terms of the nature of the language provided, and especially how the LFP elicits knowledge of

function words and syntax. We might now draw four potential conclusions about Lex30 that relate to these differences between the LFP and Lex30. First, with Lex30 subjects are not required to demonstrate knowledge of a wider context. Second, with Lex30 subjects are not expected to fit the items they provide within a wider context. Third, with the LFP subjects are not providing as many infrequent items as with Lex30, probably because they are not confident of fitting them into a wider context. Fourth, with the LFP subjects might be avoiding rarer words because they lack the knowledge of the surrounding items with which such words may fit.

The Brainstorm Frequency Profile task was thus designed to elicit a higher percentage of infrequent items than the LFP. The results show that the Brainstorm Frequency Profile did indeed successfully elicit a greater proportion of infrequent items than the LFP. However, there was still no significant correlation between Lex30 scores and Brainstorm Frequency Profile task scores.

The second issue is to examine a possible explanation for why the Lex30 scores and Brainstorm Frequency Profile task scores do not correlate. When we compare scores calculated by defining infrequent items in the same way, the Lex30 (38.5% (sd 5.9)) and Brainstorm Frequency Profile (28.1% (sd 9.1)) percentage mean scores are not vastly dissimilar. The standard deviations, however, indicate that the subjects responded to the Lex30 and BFP tasks in different ways. The lower standard deviations for the mean Lex30 scores appear to indicate that the subjects performed in a more similar way to each other than for the Brainstorm Frequency Profile scores. One potential explanation for the different scores relates to the difference in mean numbers of items produced for the two tasks. Subjects produced fewer words in response to the Brainstorm Frequency Profile task (mean words produced 51.6 (sd 20.2)) than Lex30 (mean words produced 85.2 (sd 19.5)). Thus, the mean number of words produced for the Brainstorm Frequency Profile task is lower than for Lex30. The standard deviations for the mean number of words produced appear to indicate that the subjects performed in a

proportionally more similar way to each other for Lex30 than for the Brainstorm Frequency Profile task.

An examination of the tasks and the different ways in which they attempt to elicit responses, begins to reveal why the Brainstorm Frequency Profile task scores and Lex30 scores are different. For the Brainstorm Frequency Profile task, subjects are asked to respond to one cue question, which is: *'Should a government be allowed to limit the number of children a family can have?' Write as many one-word responses as possible to this idea considering basic human rights and the danger of population explosion.* For Lex30 the subjects are asked to write down any, and up to four, words that they can think of in response to the cues provided (the first five are shown below by way of an example) in figure 6.2.

Figure 6.2 Example of Lex30 task.

Look at the words below. Next to each word, write down any other words that it makes you think of. Write down as many as you can (more than 3, if possible). It doesn't matter if the connections between the word and your words are not obvious; simply write down words as you think of them.

CUE	RESPONSES
1. attack	
2. board	
3. close	
4. cloth	
5. dig	

There are two potential reasons for Lex30 eliciting a higher mean proportion of scoring items than the Brainstorm Frequency Profile task. The first reason relates to the semantic fields activated by the two tasks. Lex30 gives a new semantic field with each of its 30 cues, while the Brainstorm Frequency Profile task provides just one semantic field. The second reason relates to motivation. Subjects might have been more inspired or

motivated to respond to the 30 Lex30 cues rather than the single Brainstorm Frequency Profile task cue, when they might have become bored or exhausted of ideas.

The third issue is to investigate whether the difference between the tasks might relate to what is being elicited. For Fitzpatrick and Meara, "The basic premise of this test [Lex30] was that a representative sample of words could be elicited from the productive L2 lexicon, using a word association task. This sample could then be categorized according to word frequency in order to measure the lexical resource of the test-taker" (2004: 55). The higher Lex30 scores (see table 6.7) and lack of correlation between the three tasks (table 6.8) show that the tasks each elicit different samples from a subject's lexicon. The lack of correlation suggests that the three tasks might not be eliciting 'a representative sample' from the subjects' lexicons. Lex30 elicited a greater number of words than the Brainstorm Frequency Profile task, and this difference indicates that the samples that the two tasks elicit cannot both be representative of the subjects' lexicons. Working on the assumption that the greater number of words elicited is likely to be the more representative, then the fact that Lex30 elicited more words than the Brainstorm Frequency Profile task suggests that Lex30 offers the most accurate representation of a subject's lexicon. What we need to know is whether the elicited vocabulary is truly representative of the subject's lexicon, which suggests that we need to determine the number of words necessary to elicit in our tasks, to be representative of our subjects' lexicons.

Thus, this third section suggests that the lack of a significant correlation between Lex30 and Brainstorm Frequency Profile task scores might be due to the difference in what is being elicited by each task. I used percentage scores because this method matched Laufer and Nation's LFP calculation. However, raw scores would reflect, in addition to the range of available vocabulary, aspects of fluency and motivation. One way to explore whether these aspects do influence scores is to calculate and compare raw scores for Lex30 and the Brainstorm Frequency Profile task. A significant correlation between Brainstorm Frequency Profile task raw scores and Lex30 raw scores should tell us that

factors besides infrequent vocabulary knowledge can influence the scores, with the difference being in what is being elicited. For Lex30 and Brainstorm Frequency Profile raw score comparison see tables 6.9 and 6.10. The two studies reported in this chapter calculate scores as percentages, the second study finding no significant correlation between Lex30% scores and Brainstorm Frequency Profile% scores. The percentage scores are calculated using the total number of words produced and show the proportion of spontaneously produced infrequent items. The second section showed that the subjects in the second study produced many more words in response to Lex30 than the Brainstorm Frequency Profile task and this disparity in the number of words produced might depend on many factors, including motivation and fluency; that is, factors other than the subjects' productive vocabulary knowledge. It seems impossible to measure the subjects' motivation objectively, but one means of determining whether factors, other than the subjects' productive vocabulary knowledge, led to their producing more words is to see whether the raw scores correlate.

Table 6.9 Lex30 and Brainstorm Frequency Profile mean raw scores, and LFP% mean scores (2k+AWL+Off-list).

	Mean score (sd)	Min score	Max score
Lex30 raw	32.8 (8)	12	48
BFP raw	14.2 (8.5)	3	43
LFP %	9.1% (4.7)	0.6%	20.7%

Table 6.10 Lex30 raw score, Brainstorm Frequency Profile raw score and LFP% score (scored using 2k+AWL+Off-list) correlations.

	LFP%	BFP raw
Lex30 Raw	0.048 p=.670	0.328 p<.001
BFP raw	0.128 p=.259	

To analyse these data, table 6.10 shows the correlations between these Lex30 raw, Brainstorm Frequency Profile task raw, and LFP% scores. There is a significant correlation between Lex30 raw scores and the Brainstorm Frequency Profile task raw scores offering some indication that raw scores are influenced by factors other than the proportion of infrequent vocabulary subjects produce. This third section shows that the raw score is influenced by factors besides productive vocabulary knowledge, such as motivation and fluency. Despite the lack of correlation between percentage scores, we might nevertheless conclude that the percentage scores are more suited to our objectives because these scores are influenced by productive vocabulary knowledge, as opposed to the additional factors that influence a raw score.

The fourth issue is to investigate potential reasons for the differences between the scores in the two studies reported in this chapter compared to earlier studies (see table 6.11). The Lex30 tasks were scored using the WebVP for both studies, one and two, and these scores were higher than the Lex30 scores recorded in earlier chapters, and also those scores scored using the JACET8000 (Jacet 2003) word lists and the word lists used by the WebVP (the General Service Lists (1k and 2k) (GSL (West 1953)), the AWL, NiL). The much higher Lex30 scores from the WebVP (particularly for chapter six, study one) suggest that something different happens when we score Lex30 in this way.

Table 6.11 Comparison of Lex30 raw scores and different frequency bands from chapters 3 to 6.

Chapter	Lex30 score (sd) and Word list		
	Nation (1984)	JACET8000 (2003)	GSL1k and 2k, AWL, NiL
3	27.4 (7.3)		
4		16.3 (8.1)	
5 – test time 1		24.1 (8.6)	
5 – test time 2		28.3 (9.1)	
6 study one		35.5 (7.4)	39.1 (7.9)
6 study two		29.5 (7.9)	32.88 (8)

Table 6.11 shows that the Lex30 scores reported in the two studies in chapter six are higher than those scores of the studies reported in the previous chapters and this phenomenon needs exploring. The major difference between scoring procedures in those studies is the use of Cobb's online WebVP scorer in chapter six. I used Cobb's (Cobb 2010) online scorer in order to retain the same scoring system as Laufer and Nation's (1995) LFP, which used the same (The General Service List (GSL) 1k, GSL 2k, AWL, NiL)) word lists. An examination of the scoring systems shows that Cobb's scorer profiles words in a different way compared to that of the JACET word lists. The WebVP uses the General Service List for the first two thousand words (1k and 2k), and while Cobb classifies 824 items from the JACET first thousand word list as 1k items, he classifies 978 items from the General Service List (GSL) (West 1953) as 1k items. This difference appears to stem from the contrasting ways in which the word lists were compiled. The JACET8000 word lists were ranked according to the frequency with which Japanese learners encounter English words (Ishikawa and Uemura, 2004: 333-347) (see Appendix 6 for a list of the JACET8000 1k words). The GSL, by contrast, was used originally as a resource for "compiling simplified reading texts into stages or steps"

(Schmitt and McCarthy, 1997: 14). The GSL compiles word families, each with its own frequency (e.g. *forgets, forgot, forgetting*), all under one head word (e.g. *forget*), However, the GSL is based on somewhat antiquated frequency studies meaning that the GSL 1k list contains both some unexpected and some items that we might these days classify as frequent words might appear further down the list words (see Appendix 7 for a list of the GSL 1k words). This difference in methods of compilation and classification of frequent items might explain why the WebVP produced higher scores than the JACET lists, and the study in chapter three (which used Nation's (1984) word lists, Cobb's WebVP 1k comprises only 600 of Nation's 1k).

6.4 Conclusion.

The studies in this chapter compared Lex30 with two tasks designed to elicit freely selected items. The first study compared Lex30 with the Lexical Frequency Profile. The results from this first study generated a lack of significant correlation between Lex30 and the LFP task. The discussion of the results suggested three reasons for the lack of a significant correlation. First, the subjects might have been limited by the LFP topic. Second, subjects need to include many function words in their LFP compositions. Third, the scoring systems for Lex30 and LFP define infrequent items differently. In order to address some of these issues the second study introduced the Brainstorm Frequency Profile task, a non-discursive version of the LFP.

The results from the comparison of the three tasks showed no correlations, even after altering the scoring for the Brainstorm Frequency Profile task and the LFP to define infrequent items in the same way as Lex30 (i.e., to include 2k items). The lack of correlation between Lex30 and the Brainstorm Frequency Profile task might relate to the disparate number of semantic fields (30 and 1, respectively), or might relate to motivation. In this second sense, subjects may have become bored when responding to the single Brainstorm Frequency Profile task cue. Given that the three tasks elicited such varied numbers of words, we still need to find a task that compares with Lex30 in terms

of eliciting a representative sample from our subjects' lexicons. We also saw that because raw scores are influenced by factors besides productive vocabulary (such as motivation and fluency) percentage scores seem better suited to our objective of measuring only infrequent productive vocabulary items. The higher Lex30 scores reported in chapter six are probably due to the GSL word list being used as the basis for the WebVP scorer. The somewhat dated GSL classification of infrequent items appears incompatible with the JACET word lists (used in chapters 4 to 6) and Nation's (1984) word lists (used in chapter 3).

In tentatively working towards a better understanding of precisely what it is that Lex30 measures, this chapter sought to further test the validity of Lex30 by attempting to compare it with a similar test. The comparisons between Lex30 and the Brainstorm Frequency Profile percentage and raw scores indicate that the key difference might lie in what is being elicited. This non-discursive version of the LFP (the Brainstorm Frequency Profile task) elicited a greater proportion of infrequent items than the standard LFP. The Brainstorm Frequency Profile task cue was able to elicit spontaneously produced infrequent vocabulary items, although not really enough of them to compare with Lex30. These findings appear to suggest that if we increase the number of cues (based on the small number of items the Brainstorm Frequency Profile task elicited probably due to its singular semantic field) and maintain the non-discursive element of the Brainstorm Frequency Profile task we might then elicit a greater proportion of infrequent items with a 'similar test'. Accordingly, chapter seven introduces a task, the GapFill task, designed to incorporate these key criteria in order to be genuinely comparable to Lex30.

Chapter 7 Comparing performance on Lex30 with performance on the Productive Levels Test and a GapFill task

7.1 Introduction.

The comparisons in chapter six showed no correlation between Lex30 and the LFP or the Brainstorm Frequency Profile. In the discussion (6.3), I suggested that this might be because the LFP elicited function words and syntax. The Brainstorm Frequency Profile elicited a greater proportion of infrequent items than the standard LFP, but a smaller proportion of infrequent items than Lex30. As the Brainstorm Frequency Profile task only has one semantic stimulus, while Lex30 provides a new semantic stimulus for each cue, this might have contributed to the lack of significant correlation between the two tasks. One way of exploring this possibility further is to test Lex30 alongside a task that demands access to a similar number of semantic fields and has similar features to Lex30 (is not discursive and uses frequency bands to measure vocabulary knowledge). Such a task might elicit a greater proportion of infrequent items. In this study, then, I introduce a GapFill task designed specifically for this purpose. Like Lex30, the GapFill task has no pre-determined answers, and each question is designed to elicit a number of items (though in a slightly more constrained way than Lex30). Given that the format of the GapFill task is, superficially at least, similar to the Productive Levels Test I use that test as a control test in this study.

In her comparisons between the Productive Levels Test and Lex30, Fitzpatrick (2007) found relatively weak but significant correlations (0.504 $p<.01$) prompting her to suggest that although "the tests claim(s) to test productive vocabulary, in fact they (all) test different aspects of this" (2007: 129). Fitzpatrick proposes that the relatively weak correlations between Lex30 and the Productive Levels Test is due to the Productive Levels Test requiring subjects to respond with a set of pre-determined words.

This chapter introduces a GapFill task with a similar number of cues to Lex30, each activating a new semantic field designed to elicit the same number of items as Lex30. A

proportion of these items will potentially be infrequent and, like Lex30, elicited by using multiple cues. The GapFill task provides context for items as does the Productive Levels Test. The GapFill task requires subjects to respond with up to five words for each of its twenty-four prompt sentences (5x24=120 words, and is designed to elicit the same theoretical maximum number of words that Lex30 elicits: 4x30=120 words). The aim of the GapFill task is to elicit a range of possible answers in the form of single-word responses to a sentence completion task. For example, the first of the 24 sentences in the GapFill task asks subjects to complete "She loved to _____ over the phone". Possible answers could therefore include *conversation, talk, chat, discuss,* and so on. The GapFill task requires up to five alternative words for each gap, and any word provided is accepted as long as it is spelled approximately, as is the case with Lex30.

The criteria required in a collateral test are that it is designed to: elicit a similar number of responses to Lex30, offer a similar number of semantic cues to Lex30, and avoid the influence of factors, other than knowledge of infrequent vocabulary items (such as motivation or fluency), that might skew the test scores. The GapFill task, I believe, meets these criteria very well. The Productive Levels Test partly meets these criteria because it elicits 90 responses, with 90 semantic cues. However, the problem with the Productive Levels Test, is that it requires knowledge of only one possible item per question, which must conform to the meaning, semantics, and orthography of each test sentence. The design of the GapFill task sought to address this issue and so words provided in response to the GapFill task do not have to fit semantically or morphologically and are accepted if approximately spelled. In this way, the GapFill task gives credit for knowledge of any infrequent words, not predetermined ones, and targeted threshold knowledge like Lex30, and unlike the Productive Levels Test.

In the design of the GapFill task, I wanted to compare the GapFill results as closely as possible with the Lex30 results. Since Lex30 does not reject responses if they are morphologically or semantically inaccurate, I also chose not to reject such responses to the GapFill task for the same reason. In terms of the two collateral tests' 'activation

properties,' Fitzpatrick (2007) suggests that the Productive Levels Test activates knowledge through form as well as meaning because the start of the word is given in a particular context. With the GapFill task, there is a semantic stimulus (as with the Productive Levels Test and Lex30) but no form stimulus since subjects are able to respond with any word, not just one particular word as required by the Productive Levels Test. In summary, the GapFill task and Lex30 share the following characteristics:

- they elicit the same theoretical maximum number of words (120)
- they are designed to elicit a range of responses
- they accept any word produced (as long as it is spelled approximately well enough to be understood)
- they have semantic stimuli
- they are not designed to elicit a pre-determined word
- one mark is awarded for every infrequent word produced.

The aim of the current study is to determine whether the GapFill task will elicit a sample of subjects' lexicons to compare with Lex30 and to compare those results with those of the Productive Levels Test. Thus, the study compares subjects' performances on three tasks: the GapFill task, Lex30 and the Productive Levels Test. Accordingly, the research question is:

What is the relationship between GapFill scores and Lex30 scores, compared to a task designed to elicit pre-determined productive vocabulary (The Productive Levels Test)?

The experiment in this chapter reports on a small-scale study with only five subjects. My aim in testing with such a small subject group is to present a detailed analysis of the subjects' responses to each of the tasks. This analysis will allow me to examine the differences between the individuals' performances on each of the tasks.

7.2 Study.

7.2.1 Subjects.

The subjects were five female Osaka University L1 Japanese students, aged between twenty and twenty five, and selected from different L2 proficiency levels, ranging (in ability) from post elementary to advanced level. The subject with the most advanced level of English was a post-graduate student of Linguistics, while the subject with the lowest English proficiency level, that of post-elementary English, was an undergraduate Engineering student with a TOEFL score of 350. The three other subjects' English proficiency ranged from pre to post intermediate levels. At the time of the experiment, the students were not taking any English classes.

7.2.2 Method.

Testing took place over two different sessions, a week apart. In the first session, the five subjects took the Lex30 task, and in the same way as the Lex30 pilot (Meara and Fitzpatrick, 2000), subjects were given 15 minutes to complete the Lex30 task. Following the Lex30 task, a week later, subjects completed the remaining two tasks, the Productive Levels Test, and the GapFill task, which they took in that order.

7.2.2.1 The GapFill task.

This section describes the GapFill task, starting with the process involved in the design of the GapFill task sentences. To maximise comparability with Lex30, I needed the task to elicit a maximum of 120 potential answers, so I started with twelve sentences designed to elicit 10 words each. If I had tested with twelve sentences, I would have required ten word responses to the 12 cues. I expected that this would be demanding of the subjects, especially considering the danger (discussed in section 6.3) that fewer cues might elicit fewer responses, as was the case with responses to the single Brainstorm Frequency Profile cue. Pilot testing with non-native and native speaker subjects

indicated that five responses per subject was an optimal number, so I decided on twenty-four sentences each eliciting up to five words as this would give a maximum of 120 responses, as with Lex30. I decided to test subjects with these test sentences requiring up to five items in order to match Lex30 as closely as possible.

The twenty-four GapFill task sentences were developed according to specific principles, to elicit varied responses from the subjects, I made the decision to elicit an equal number of nouns, verbs, and adjectives on the assumption that this would generate a suitably varied set of responses. I wanted to elicit a variety of responses rather than elicit only typical responses and to avoid lexical sets (such as the Lex30 cue *furniture* eliciting the lexical set: *chair, desk, sofa, bookcase,* etc). The twenty-four sentences were therefore designed in order to meet the six following criteria, they: i) were syntactically simple; ii) did not elicit lexical sets (e.g. *banana, apple, orange,* etc.); iii) could readily elicit five responses from native speakers or proficient non-native speakers; iv) contained only high frequency words or were likely to be known by students of intermediate level or above; v) did not elicit similar words to another sentence in the task; and, vi) did not elicit single gap responses that might favour a particular response over another. Both native and non-native speaker groups piloted the GapFill sentences, thus enabling me to reject sentences that did not work. I rejected sentences if they elicited either too few responses or only highly frequent responses. Figure 7.1 shows an example of a GapFill task completed by the advanced level English proficiency subject, Hiroko:

Figure 7.1 Example of completed GapFill task.

In the spaces provided below write as many one-word responses as possible (up to five) to complete each sentence. Try not to repeat words you have already used.

1. She loved to _____ over the phone.	*talk*	*speak*			
2. When I feel sad I always go to the_____.	*hospital*	*cafe*	*room*	*bedroom*	*garden*
3. They think car-racing is_____.	*dangerous*	*exciting*	*fantastic*	*awesome*	*stupid*
4. His colleague wanted to _____ the report.	*finish*	*write*	*do*	*refuse*	
5. My favourite _____ is football.	*sport*	*thing*	*activity*	*one*	
6. She looked _____ when she saw her friends.	*upset*	*mad*	*confused*	*happy*	*sad*
7. He couldn't _____ the car.	*get*	*have*	*buy*	*sell*	*own*
8. If there was a fire in my house I would save my _____.	*parents*	*kids*	*dog*	*cat*	*girl*
9. Many people feel _ _about the environment.	*worried*	*optimistic*	*pessimistic*	*nothing*	
10. The parents _____ the children.	*love*	*educate*	*discipline*	*hate*	*raise*
11. He was happy with his _____.	*children*	*result*	*success*	*money*	*girlfriend*
12. He didn't think her teacher was _____ at all.	*intelligent*	*stupid*	*mean*	*beloved*	*weird*
13. She always wanted to _____ after a busy day at work.	*sleep*	*drink*	*swim*	*run*	*smoke*
14. She sent _____to her mother.	*it*	*cards*	*presents*	*tickets*	*food*
15. The weather looked ____ before the game.	*great*	*nasty*	*awful*	*beautiful*	*fine*

16. He wanted to _____ the letter.	*send*	*receive*	*hide*	*keep*	*rewrite*
17. She was excited about _____.	*travelling*	*skiing*	*skating*	*diving*	*camping*
18. The girls thought the rock concert was _____.	*amazing*	*super*	*incredible*	*gorgeous*	*excellent*
19. He took the chance to _____ the president.	*become*	*be*	*meet*	*see*	
20. He gave his boss _____.	*results*	*copies*	*agenda*	*coffee*	
21. At the funeral the family felt ___.	*disappointed*	*sick*			
22. He always _____his breakfast.	*eats*	*misses*	*brings*	*likes*	
23. She put the food in the _____.	*fridge*	*basket*	*microwave*	*box*	*freezer*
24. She was always _____ to those who needed help.	*kind*	*smiling*	*friendly*	*unfriendly*	*gentle*

The subjects had fifteen minutes to complete the GapFill task, the same time limit as the Lex30 task, because it aimed to elicit the same number of responses potentially as the Lex30 task. Scoring of the GapFill task responses was also the same as with Lex30 data, hence I accepted any response from the subjects, even if it was considered ungrammatical, badly spelled, or inappropriate. The only criterion for rejecting an item was if it was illegible. This rationale is the same as that of Lex30, in the sense that any item a subject provides is accepted as long as it is spelled approximately. As in the standard Lex30 scoring, repeated words and proper names were not scored, and items were lemmatised. Each subject's GapFill score was taken to be the total number of infrequent (non 1k) words produced. I used the WebVP (http://www.lextutor.ca/vp/eng) to calculate the number of words outside the first 1000.

In order to compare GapFill scores with Lex30 in the study, I calculated a GapFill percentage score, for the reasons outlined in the discussion in section 6.3. Percentage

scores were calculated primarily to show the proportion of spontaneously produced infrequent items. The GapFill percentage score was each subject's GapFill raw score divided by the total number of items produced, and then multiplied by 100.

7.2.2.2 Lex30.

The scoring for Lex30 was conducted in the same way as reported in chapter six, and similar to the standard (Meara and Fitzpatrick 2000), with the only difference being that the scores were processed with Cobb's WebVP (http://www.lextutor.ca/vp/eng/). Each subject wrote their responses to Lex30, and their paper tests were then typed into a computer text file. Repeated words and proper names were not scored, and items were lemmatised. Each subject's Lex30 data was then scored using the online WebVP. Section 6.2.1.3 provides a detailed account of how the WebVP was used.

7.2.2.3 The Productive Levels Test.

The Productive Levels Test (see section 2.2.2) was presented in the same way as by Schmitt, Schmitt and Clapham (2001). Appendix 5 shows the Productive Levels Test used in this experiment, in which 18 pre-selected items are tested at each level (5 levels x 18=90). Below is an example of a test item at the 5000 word level to elicit *oath*.

1. Soldiers usually swear an oa___ of loyalty to their country

In the same way as Laufer and Nation (1999), I did not penalize subjects for minor spelling mistakes. In addition, as in Laufer and Nation (1999) I interpreted the Productive Levels Test score as being the total number of correct responses a subject produced.

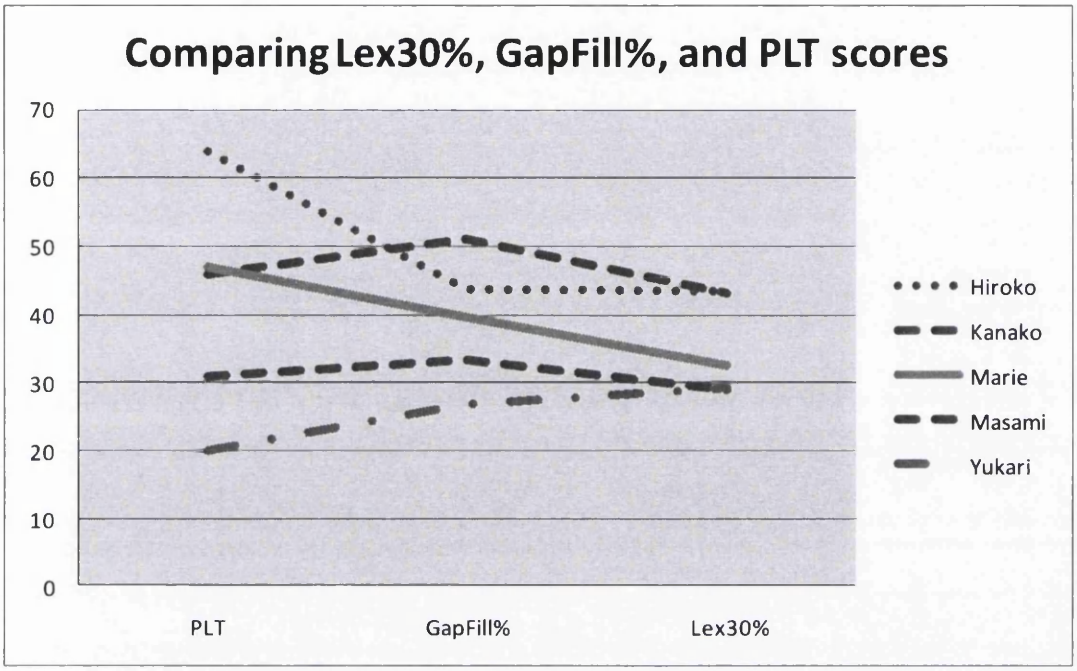
7.2.2.4 Results.

Table 7.1 shows the scores for the three tests and figure 7.2 shows the results in graph form. Although the Productive Levels Test scores are not really directly comparable with the other scores, as it uses a different scale, figure 7.2 shows that there is minimal overlap of the lines and also shows that Yukari is the lowest scoring subject, then Masami, then Marie, and then Hiroko, with Kanako as the highest scoring subject. GapFill% scores are higher than the same subjects' Lex30% scores for all but Yukari. The ranking of the GapFill% scores appears broadly predictive of the subjects' Lex30% scores. I have not included a correlation analysis because the number of subjects (5) is so small.

Table 7.1 Lex30%, GapFill% task, and Productive Levels Test Scores.

	Lex30 %	GapFill %	Productive Levels Test
Hiroko	43.8	43.6	64
Kanako	43.6	51.1	46
Marie	32.4	39.6	47
Masami	29.1	33.3	31
Yukari	29.3	26.9	20

Figure 7.2 Comparing Lex30%, GapFill% task, and Productive Levels Test scores.



7.3 Discussion.

The research question for chapter seven asked 'What is the relationship between GapFill scores and Lex30 scores, compared to a task designed to elicit pre-determined productive vocabulary (The Productive Levels Test)?' As the Productive Levels Test is an established testing tool and appears to have proven validity, I wanted to compare its scores to those of Lex30. The design of the GapFill task attempted to exploit the positive features of both the Productive Levels Test and Lex30, based on the results from chapter six that showed a lack of correlation between Lex30 scores and LFP scores or the Brainstorm Frequency Profile scores. The GapFill task was therefore devised to elicit a similar number of responses to Lex30, using a similar number of semantic fields to Lex30, and to avoid the interference of factors other than knowledge of infrequent vocabulary items that might skew the test scores. In short, the experiment compared GapFill scores with Lex30 scores and with Productive Levels Test scores with the aim of examining the differences between the scores produced by the three tasks.

The results appear to indicate that there is a closer relationship between the scores produced by these three tasks than between the three tasks compared in chapter six (Lex30, the LFP, and the Brainstorm Frequency Profile). In the discussion that follows, I examine why this is the case and why GapFill scores seem to be a good indicator of Lex30 scores. I have divided the discussion into five sections. The first section examines the relevance of the GapFill's similar number of semantic cues to Lex30. The second section examines different response behaviours to the Productive Levels Test and Lex30. The third section explores the activation properties of the tasks, then discusses the different aspects of knowledge measured and addresses the differences between eliciting and scoring. The fourth section examines the implications of the non-predetermined items elicited by the GapFill task, as with Lex30, compared to the Productive Levels Test. The fifth section discusses the relative advantages of Lex30 and the GapFill tasks. The sixth section examines a potential benefit of the GapFill task, relating to the enhanced ability to provide feedback to test-takers.

The first section examines the relevance of the GapFill's similar number of semantic cues to Lex30. Both tests elicit freely generated vocabulary using a similar number of cues, or semantic fields. The results from this study are encouraging because they appear to suggest that when we compare Lex30 task scores with GapFill task scores similar proportions of infrequent vocabulary items are produced. As we saw in the results section, the GapFill% task elicits a greater proportion of infrequent items than the Lex30% task for four of the five subjects (only Yukari's Lex30% scores, the lowest scoring subject across all tasks, were higher than her GapFill% scores). These similar Lex30 and GapFill task scores appear to demonstrate that the two tasks elicit similar proportions of infrequent items. The twenty-four GapFill cues appear to have successfully elicited similar proportions of infrequent items to those elicited by Lex30 (see tables 7.2 and 7.3); however, the GapFill task elicited a mean number of 90 (sd 13.2) total responses from the subjects, compared to a mean number of 108 (sd 10.3) total responses for Lex30.

The second section examines the differences manifest when Lex30 response behaviour is compared to Productive Levels Test response behaviour. A comparison of subjects' scores and performances on the Productive Levels Test and Lex30 makes these differences clear. The Productive Levels Test scores are more varied than the Lex30 and GapFill task scores. For example, Hiroko and Kanako were the two top scoring subjects on the three tasks; however, the fact that Hiroko scored much more highly than Kanako on the Productive Levels Test is somewhat unexpected in light of their broadly similar Lex30 and GapFill scores. One potential difference between the two tasks might lie in the particular response behaviour exhibited by particular subjects. For instance, Kanako produced fewer responses than Hiroko when responding to the Productive Levels Task, possibly intimating that Kanako may only have written items of which she felt confident. Conversely, Hiroko took more risks than Kanako in providing a greater number of responses to the Productive Levels Test. Although many of Hiroko's responses were written inaccurately, they nevertheless constituted scoring items. We might also speculate why one subject might perform better at certain levels of the Productive Levels Test than another subject, based on a breakdown of Lex30 scores. Table 7.2 shows a breakdown of Hiroko and Kanako's Lex30 scores: Hiroko's Lex30 score consists of a greater number of AWL and Off-List items, whereas Kanako's Lex30 score consists of a greater number of 2k items. Similarly, table 7.3 shows a breakdown of Hiroko and Kanako's Productive Levels Test scores, with Hiroko's score consisting of more 10k (10) items than Kanako's (2). The breakdown of the two subjects' respective Lex30 and Productive Levels Test scores reveals that Hiroko appears to have a better knowledge of less frequent (than 2k) items than Kanako, possibly offering an explanation for why Hiroko fared better at the less frequent levels of the Productive Levels Test. However, as well as being potentially useful for revealing such common tendencies across both tests, this kind of analysis also exposes a potential weakness with Lex30's scoring system in that it does not make any distinctions between those subjects who know a large number of 2k words and those who know a large number of more infrequent (e.g. 10k) items.

Table 7.2 Two subjects' Lex30 responses.

	1k	2k	AWL	Off-list	Lex30 raw score	Lex30 % score
Hiroko	59	26	9	11	46	43.8
Kanako	62	37	3	8	48	43.6

Table 7.3 Two subjects' Productive Levels Test responses.

	2k	3k	5k	UWL	10k	PLT score
Hiroko	17	15	10	12	10	64
Kanako	15	10	7	12	2	46

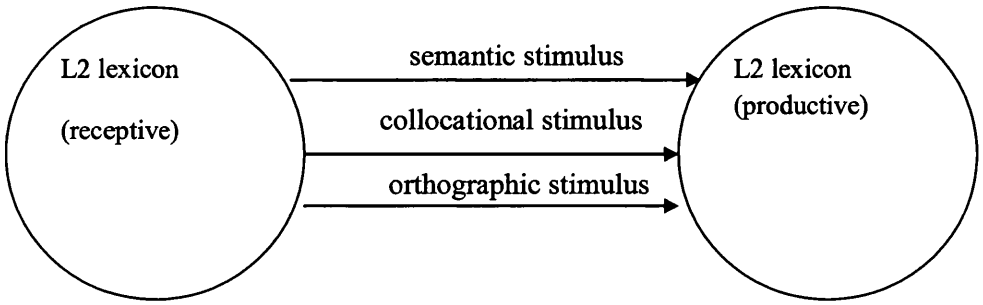
However, we cannot compare the scores solely in terms of the proportion of infrequent items produced, because the scoring for the Productive Levels Test is different to the scoring for Lex30 and the GapFill tasks, we need an alternative means to explore the differences between scores on the three tasks. Fitzpatrick (2007) suggests that we should examine whether tasks of productive vocabulary knowledge elicit "information about the same kind of vocabulary knowledge in the same way" (p. 127) in the form of two approaches. The first approach relates to the different 'activation properties' for each task (Fitzpatrick 2007: 127). The second approach relates to the different aspects of knowledge (Nation 1990) that are addressed by different tests. The following two sections of the discussion address each of these two approaches in turn.

The third section of the discussion explores the activation properties of the tasks by representing each test in terms of their different models of activation with each test having different 'activation properties'. Figure 7.3 below shows the models of activation for the three tasks. Fitzpatrick suggests that Lex30 "has only one activation property: the

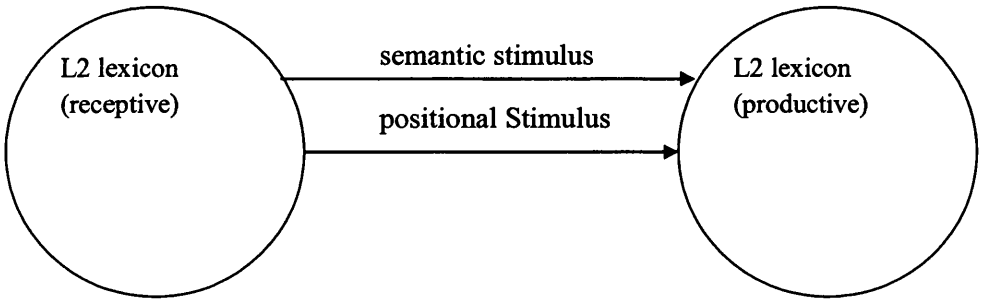
L2 semantic stimulus" (2007: 127). The GapFill task has the same activation property as Lex30, the L2 semantic stimulus, but also has an additional activation property: a positional stimulus. Meanwhile, the Productive Levels Test "has three activation properties: L2 semantic stimulus, L2 orthographic stimulus and an L2 collocational stimulus" (Fitzpatrick 2007: 127). Here, I am using the terms 'collocational' and 'positional' interchangeably. This breakdown appears to indicate that the Productive Levels Test utilizes a greater number of activation properties than Lex30 and the GapFill task, and shows that the three tasks elicit different information about different kinds of vocabulary knowledge in different ways. It is worth exploring, briefly, why the GapFill task elicits fewer items in total than the Lex30 task. The subjects appear to have edited out responses that they did not consider grammatically or semantically possible. Although the subjects were not scored for the relative appropriateness of their responses, they nevertheless appear to have considered whether a response was appropriate or not when they took the GapFill task which suggests an additional activation property when compared to Lex30.

Figure 7.3 Models of activation for the three tests (adapted from Fitzpatrick 2007: 128).

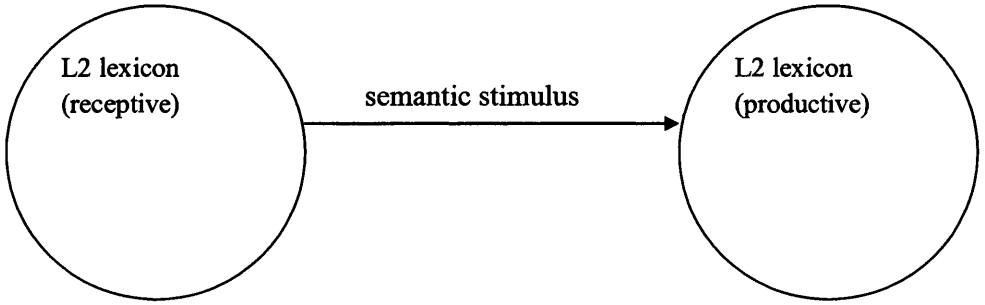
Productive Levels Test



The GapFill task



Lex30



Here, I discuss the different aspects of knowledge measured on the three tasks, and address the differences between the eliciting and scoring of each. Nation's (1990) table categorizes tasks according to the type of knowledge demanded: receptive or productive, form, position, function, and meaning knowledge. Table 7.4 below shows a revised version Fitzpatrick's adaptation of Nation (2007: 129) to include the GapFill task. Fitzpatrick (2007) suggests that Nation's table is useful because it offers a framework in which to identify the multiple aspects of vocabulary knowledge (including receptive knowledge) that are being accessed. While Lex30 and the GapFill task only appear to access three of Nation's categories, the Productive Levels Test accesses six of Nation's categories. Yet the different performances on the tasks appear dependent on the different aspects of knowledge that are elicited. It is quite challenging to categorise these aspects of knowledge confidently because of the probable variation within particular ticked boxes. For instance, a tick for 'What word should be used to express this meaning?' for both the GapFill task and Lex30 implies only threshold knowledge since we do not know whether subjects can express the meaning; rather, their understanding is implied. For the Productive Levels Test subjects have to demonstrate explicitly that they understand the meaning of the word because of the comparatively strict constraints of the sentence completion task. The differences in the subjects' performances on the tasks also appear to be dependent on the differences in scoring. For Lex30 and the GapFill tasks, subjects are required to produce any items they know in response to the cues. The subjects then receive credit for any items produced outside the first thousand frequency band. The demands of the GapFill and Lex30 tasks remain constant throughout the 15-minute tasks in the sense that the difficulty of the task does not change. Yet by contrast, the Productive Levels Test, becomes more demanding as subjects progress, and this occurs in two ways. Firstly, the surrounding words that constitute the test sentences include increasingly more infrequent items. Secondly, the pre-determined words the test aims to elicit become increasingly more infrequent. The second of these two test demands is discussed in detail in the third section of the discussion below. This section of the discussion suggests that the Productive Levels Test accesses a greater number of

aspects of knowledge than both the GapFill and Lex30 tasks. In addition, the scoring difficulty increases in the Productive Levels Test as subjects advance through the task whereas the scoring difficulty remains constant throughout the Lex30 and GapFill task.

Table 7.4 Aspects of word knowledge (adapted from Fitzpatrick 2007 and Nation 1990) tested by Lex30, the GapFill task (GF), the Productive Levels Test.

Aspect of Word knowledge (R=receptive, P=productive)			Lex30	GF	PLT
Form: Spoken form	R	What does the word sound like?			
	P	How is the word pronounced?			
Form: Written form	R	What does the word look like?			
	P	How is the word written and spelled?	✓	✓	✓
Position: grammatical position	R	In what patterns does the word occur?			✓
	P	In what patterns must we use the word?			✓
Position: collocations:	R	What words or types of words can be expected before or after the word?		(✓)	✓
	P	What words or types of words must we use with this word?			
Function: frequency	R	How common is the word?			
	P	How often should the word be used?			
Function: appropriateness	R	Where would we expect to meet this word?			
	P	Where can this word be used?			✓
Meaning: concept	R	What does the word mean?			
	P	What word should be used to express this meaning?	✓	✓	✓
Meaning: associations	R	What other words does this word make us think of?			
	P	What other words would we use instead of this one?	✓	✓	

In this fourth section, I discuss the implications of not-predetermining items to be elicited, as is the case with the GapFill task and Lex30, compared to the Productive Levels Test. Both the GapFill task and the Lex30 task give credit for providing any

approximately spelled infrequent items outside the first thousand frequency band. The Productive Levels Test, by contrast, gives credit in a much narrower way by requiring that subjects provide a particular, pre-determined word, pre-selected as being representative of each particular frequency band, with the words escalating in infrequency for each band tested. For the Productive Levels Test, subjects are either correct or not in response to each of the 18 sentence completion tasks chosen to test for knowledge of each band. For both the GapFill task and Lex30, subjects have a greater chance of success, in the sense that they have greater flexibility and thus a greater number of opportunities (i.e., more than one) to provide a scoring item or scoring items in response to each cue. The GapFill task provides five such opportunities per cue to provide a scoring item, while Lex30 provides four. Thus, the main difference between the three tasks appears to be that providing a scoring item in response to the Productive Levels Task requires that subjects respond to a much tighter set of demands, than the GapFill and Lex30 tasks, because the Productive Levels Test scores pre-determined items.

In this fifth section, I discuss two broad relative advantages of using the GapFill task and the Lex30 task. Firstly, as the first section of this discussion highlighted, a greater number of words were generated by the Lex30 task than the GapFill task. As discussed above, a potential reason for the subjects providing fewer words in response to the GapFill task was that they might have edited out responses that they did not consider grammatically or collocationally possible. As figure 7.3 shows, the additional 'positional' activation stimulus for the GapFill task appears to influence subjects' responses. While subjects were not scored for their ability to provide grammatically or collocationally possible responses, they do nonetheless appear to have made use of these aspects of knowledge when responding to the GapFill task. Secondly, as both tasks have a small number of activation properties that access the same aspects of knowledge this offers plausible support for the construct validity of Lex30 as a test of 'productive vocabulary'. The results from this experiment offer encouraging support for further

experimentation with a much larger subject group to compare Lex30 and GapFill test scores to determine whether such a correlation exists between the two tasks in terms of measuring productive vocabulary knowledge.

Lastly, there is one further potential benefit to using the GapFill task, one which should be acknowledged as it relates to the practical use of the test. As the subjects appear to have provided only grammatically or collocationally possible responses to the GapFill task, one potential benefit of the task is the facility to provide feedback to test-takers. Comparing and evaluating class answers to the GapFill task could be an excellent vocabulary building exercise, and could facilitate useful discussion of what makes some words more fitting than others.

7.4 Conclusion.

This chapter compared performance on Lex30 with performances on the Productive Levels Test and a GapFill task. The results showed a close relationship exists between the three tasks, but particularly between Lex30 and the GapFill task. I discussed five issues. The first related to the similar number of semantic fields for the GapFill task and the Lex30 task as having the potential to account for the similar scores. The second related to the different kinds of response behaviour we examine when we analyse the test scores closely. This second suggestion raised a particular drawback with Lex30 in that it doesn't distinguish between subjects who know many 2k words compared to those who know more 10k words. The third related to the additional positional stimulus influencing production and the scoring on the GapFill task as opposed to the Lex30 task. The fourth related to the different demands that each test makes on the subjects as accounting for the different scores. The fifth suggested that the similarity of the Lex30 and GapFill tests scores appears to offer plausible support for the construct validity of Lex30 as a test of 'productive vocabulary,' and encourages further experimentation with a much larger group in order to compare Lex30 and GapFill test scores more comprehensively.

The results from this chapter tentatively suggest that I might claim to have identified the constructs of Lex30 to a point where I can design a similar test, and future research will see whether this new test can improve on what Lex30 has to offer. Finally, I suggested that the GapFill task could also serve as an excellent teaching resource for vocabulary building if subsequently used for discussion in class.

This chapter set out to compare Lex30 with the GapFill task and the Productive Levels Test, in order to enhance understanding of exactly which aspects of productive vocabulary are being measured in the Lex30 test. The tests examined in chapters six and seven all superficially test productive vocabulary but produce different results and activate vocabulary in different ways. The comparisons between Lex30 and the GapFill task suggest that, when we compare Lex30 with a task that has a similar number of semantic fields and activation properties that activate the same aspects of knowledge we generate very similar test scores. These similarities and the reasons underlying them are discussed at length in the next chapter.

Chapter 8 Discussion

8.1 Introduction.

The experimental chapters attempted to examine what Lex30 measures in terms of the construct of productive vocabulary. We might tentatively conclude that Lex30 only accesses minimal aspects of productive vocabulary knowledge, which appears to support the concerns presented at the end of chapter two (see table 2.8). Table 2.8 shows a comparison of tests of productive vocabulary, highlighting that Lex30 potentially only accesses knowledge of the form and meaning of a word. In this chapter, and based on the findings from my experimental chapters, I consider a number of potential issues that pertain to Lex30 as a test of productive vocabulary.

The discussion that follows is divided into five sections. The first section (8.2) addresses three concerns that relate to the measuring system used. I first examine the raw and percentage scoring systems to determine which is the more appropriate when attempting to measure productive vocabulary. Second, I discuss how useful the various frequency lists on which Lex30 bases its scoring are in providing an estimate of subjects' productive vocabulary knowledge. Third, I examine how effective Lex30 is at sampling the contents of the lexicon. The second section (8.3) deals with three concerns that relate more broadly to the kind of knowledge that Lex30 is measuring. First, I examine attempts to describe vocabulary knowledge as part of general knowledge. Second, I examine wider vocabulary knowledge, and the aspects of vocabulary knowledge measured by Lex30. Third, I examine comparisons with other tests of productive vocabulary and consider the various interpretations of the construct of productive vocabulary knowledge. The third section, (8.4) examines lexical processing and Lex30. The fourth section (8.5) examines a bilingual model in light of Lex30 response behaviour. The fifth and final section (8.6) examines the construct of productive vocabulary and Lex30.

8.2 The knowledge Lex30 measures: Scoring systems, frequency lists, and sampling the contents of the lexicon.

8.2.1 The raw and percentage scoring systems.

In the experiments described in this thesis, two distinct scoring methods have been used. The experimental chapters (chapters 3 to 7) have used Lex30 raw scores (chapters 3-7) and Lex30 percentage scores (chapters 4 to 7). This section examines the differences between the two systems and the interpretations of them.

A short comparison of the two scoring systems, of two subjects taken from the study in chapter seven, shows the differences in scoring. Hiroko, the most proficient L2 English subject in the study reported in chapter seven, produced 106 words of which 46 were infrequent so her Lex30 raw score is 46 and her percentage score is 43.8%. Yukari, meanwhile, the least proficient L2 English subject in the study reported in chapter seven, produced 92 words, of which 27 were infrequent, so her raw score is 27, and her Lex30 percentage score is 29.3%. The raw score system shows the number of infrequent items produced, while the percentage score system reflects both Hiroko and Yukari's scores as a proportion of the total number of words they produce. Hiroko produced a greater number of words overall (106) as well as providing a greater number of infrequent words (46), and this leads us to ask why one subject provides more words in response to Lex30 than another.

Here I examine the factors that might lead to one subject producing a greater number of words than another subject. The following four of which might be contributing factors:

- Relative size of lexical resource – The lexical pools of test takers, from which they select productive vocabulary items, differ in size.
- Motivation –Some subjects might feel more motivated by the Lex30 task than others.

- Writing speed – Subjects with a faster writing speed might be able to provide a greater number of words in response to Lex30 than those with a slower writing speed, and subjects with a different L1 orthography might take longer than those with the same L1 orthography.

- Fluency – Fluency influences the speed and strength of connections between cues and responses. Bilingual memory model research suggests (e.g. Potter *et al.* 1984) that with increasing expertise in the L2, or L2 fluency, individuals are better able to make a greater number of connections between their L2 words.

The raw scoring system might then provide us with information related to the above four factors. As we saw in chapter six (table 6.10), factors other than the subjects' productive vocabulary abilities influence the raw scores. However, we are unable to determine the extent to which these factors influence the Lex30 task score. The difficulty is that all four of the factors above, and potentially more, might influence the Lex30 score and we cannot tease them apart to see what is contributing to the score.

Each Lex30 raw score is a tally of all of the infrequent types that subjects produce with 'infrequent' taken to mean all of the words produced beyond the 1000 word frequency band. The raw score is affected by different aspects of language competence (including fluency, motivation, and so on, as in the bullet points above) and illustrates that we need to consider why some subjects provide more words than do others. If our purpose is to measure productive vocabulary ability then the raw scoring system is not suited to our purposes. This contrasts with the Lex30 percentage score, a score not influenced by these many aspects of language ability and only telling us about the quality of subjects' responses to Lex30 (i.e., their ability to provide infrequent items as a proportion of their Lex30 responses).

Each Lex30 percentage score is th*et al*ly of the infrequent types produced as a percentage of the total number of types produced. The Lex30 percentage score measures

only the proportion of 'spontaneously' produced words that are infrequent. The percentage score is affected by the quality (as defined by the frequency or infrequency) of words produced. Given that the Lex30 percentage scoring system is only influenced by the production of infrequent vocabulary items, it is suited to our purposes because we aim to examine the construct of productive vocabulary knowledge.

It is important to know what the scores mean, and so, for my experiments in chapter six and seven, I only used the percentage scoring method because I was only interested in the proportion of infrequent items that the subjects produced. The percentage scores indicate how often subjects produce infrequent items in relation to the other words that they provide. This section forces us to conclude that the percentage scores are more suited to our objectives.

8.2.2 How useful or fit for purpose are the frequency lists?

Lex30 scores subjects' responses using frequency lists. Subjects score a point for each infrequent response they provide, which is any word outside the first 1000 most frequent English words and up to a maximum of 120 words (30 cues x four responses= 120 words). In this section, I discuss whether the use of the frequency lists, upon which the Lex30 score is based, are fit for providing an estimate of subjects' productive vocabulary knowledge.

Testing with Lex30 assumes that the more infrequent words a learner is able to access and produce, the more advanced their lexical development is considered to be. This assumption is made because it is generally recognised that learners acquire more frequent words first (Nation, 2001: 9). Using frequency lists, therefore, assumes that learners acquire words in approximately the same order. The frequency lists are compiled from corpora, and the corpora are compiled from texts that are selected in certain ways, so the nature of the texts determines the nature of the lists. However, frequency profiling might only accurately describe particular groups of subjects whose

learning paths are reflected by particular frequency lists. Thus, the same frequency lists for one group, by contrast, might not match the learning paths of another group tested according to the same categorization of 'infrequent' words. The ways in which the lists are used in testing assume a match in terms of the order in which learners acquire words. The lists differ in terms of the way they classify words, depending on strict frequency criteria or the order in which words are encountered by particular groups of learners. Thus, although we call the lists frequency lists, some use other criteria as well.

Fitzpatrick and Meara (2004) propose using the Jacet8000 (Jacet 2003) word lists on the basis that a "more up to date set of frequency bands might improve the accuracy of the Lex30 measure" (2004:71). As the vast majority of my subjects were L1 Japanese speakers it seemed logical to use the frequency lists designed for those learners (JACET8000 (Jacet 2003)), and I did so in chapters four and five (see Appendix 6 for a list of the 1k JACET8000 words). In chapter 3, I used Nation's (1984) word list as did Meara and Fitzpatrick (2000). In chapters 6 and seven, I used the word lists used by the WebVP (GSL 1k), a description of which follows, as I needed comparability with the LFP and the Productive Levels Test, respectively. It is, therefore, perhaps important to consider how the Jacet8000 word lists were compiled. The following points summarize the features of the Jacet8000 word list.

- The JACET8000 list was compiled recently (in 2003)
- JACET8000 was designed for Japanese learners of English
- JACET8000 was based on the British National Corpus
- JACET8000 was ranked according to the frequency at which Japanese learners encounter English words.

(Ishikawa, S. and Uemura, 2004: 333-347)

The WebVP uses two lists: the GSL (the first 1000 words of the General Service List (GSL), the second 1000 words of the GSL), and the Academic Word List (AWL)

(Nation 1990). The General Service List (GSL) of English words is what Nation refers to when he writes about the first 2000 words (see Appendix 7 for a list of the 1k GSL words). The GSL is not in frequency order and is based on written texts thought to represent ideal vocabulary for L2 students. The AWL is a list of 800 vocabulary items designed for L2 students planning to study in an English language university.

One of the benefits of Lex30 is that it makes the distinction that any items produced beyond the 1k frequency band are considered infrequent which might imply only a minimal difference between scores using different frequency lists, because according to Aizawa (2006) there appears to be a lot of overlap in the first thousand frequency band. However, not all frequency lists are the same (since they depend on different corpora and different principles for the compilation as seen above and in section 6.3) and, as a further complication, Aizawa (2006) suggests that the order of vocabulary acquisition after the 4k frequency band is fairly random.

In order to compare the effects on scores of using different word lists here, I have rescored the data in chapters six and seven using the JACET8000 word lists. I have only shown raw scores in this table to show the different numbers of items identified as infrequent as judged by the two word lists respectively (and not in proportion to the number of words produced by each individual subject).

Table 8.1 A comparison of Lex30 raw scores using different frequency bands from chapters 6 and 7.

Chapter	Mean Lex30 raw score and Word list	
	JACET8000 (sd)	VocabProfile (sd)
6 study one	35.5 (7.4)	39.1 (7.9)
6 study two	29.5 (7.9)	32.8 (8)
7	34 (8.8)	38.6 (8.6)

Table 8.1 shows that the Lex30 scores are higher when we score with the lists used by the VocabProfile (which the WebVP uses), while the interval between the mean

numbers of infrequent items produced in each study is approximately the same (the difference is between 3 and 5, or by an increase of around 10%).

The profiling of responses according to JACET8000, in order to follow the learning path of the predominantly Japanese learners assessed in the experimental chapters, might provide more accurate scoring. This profiling according to particular frequency lists implies the need to match the corpora on which the frequency lists are based and the learning path of the particular learners that are the focus of the assessment. This is a potential problem for all tests using frequency lists (such as the Vocabulary Levels Test, Productive Levels Test, and the Lexical Frequency Profile) and is clearly not confined to Lex30.

This section briefly outlined some of the potential limitations of the frequency lists in use when testing with Lex30. Anybody using Lex30 would be advised to be stringent in ensuring that they choose lists relevant to the learning paths of their subjects. Lex30 is a flexible tool and can be scored using any frequency list relevant to the particular learner group.

8.2.3 How effective is Lex30 at sampling the contents of the lexicon?

Lex30 works on the assumption that the words that it elicits are a representative sample of the subject's productive lexicon. If a comparison with a similar test elicits similar test scores we might be in a position to claim that a Lex30 score might provide a broad indication, or is representative of, the number or proportion of infrequent items a subject is likely to have access to.

This section relates to the sample of words that Lex30 generates, and examines whether or not the maximum 120 words are sufficient to elicit a representative sample of the productive lexicon. This section first discusses the lowest scores elicited in the

experimental chapters and, second, discusses whether comparisons with a similar test indicate that Lex30 scores are representative.

As table 8.2 below shows, in a comparison of the Lex30 scores throughout the thesis, there is a variability in the number of words that subjects produce for Lex30. The subjects in chapter four produced the smallest mean number of words (47.5 (sd 19)) compared to the subjects in the first study in chapter six who produced the largest mean number of words (115 (sd 17.4)).

Table 8.2 Mean number of words produced and Mean Lex30 raw scores for experimental chapters 3-7.

Chapter	Mean number of words produced (sd)	Mean Lex30 raw score (sd)
3	63.4 (16.9)	27.4 (7.3)
4	47.5 (19)	16.3 (8.1)
5	110 (17)	24.1 (8.6)
6	115 (17.4)	35.5 (7.4)
7	108.6 (5.3)	38.6 (8.6)

The subject group that produced the lowest number of words also had the lowest Lex30 mean scores (chapter four Lex30 mean score 16.3 (sd 8.1)), demonstrating the significance of the difference between raw and percentage scores discussed in section 8.2.1. In chapter four, I compared Lex30 scores with independent receptive measure scores (X_Lex) and discussed whether there is a threshold number of responses below which Lex30 does not work. By cutting out lower producing subjects, I wanted to determine whether there is a better correlation between X_Lex and Lex30 with the remaining higher producing subjects (i.e. subjects who produced a greater number of items). When subjects produced 20 or more words the correlations between X_Lex and Lex30 were the strongest. This suggested that there does not appear to be a threshold number of responses *below* which Lex30 does not work, although there is a stronger and more significant correlation (with an independent receptive measure) when subjects

produce 20 or more words in response to Lex30. Despite this apparent variability in the numbers of words subjects provided in response to Lex30, and in comparison to the different experiment chapters, it appears that Lex30 appears to elicit similar *samples* from subjects' lexical stores. The question remains, however, as to whether this sample might be representative or not.

If we base our assumption that the greater number of words elicited is likely to be the more representative, then the comparisons between Lex30 and the Brainstorm Frequency Profile task in chapter six suggest that Lex30 offers the most accurate representation of a subject's lexicon. What we need to determine is whether the elicited vocabulary is truly representative of the subject's lexicon. This suggests that we need to determine the number of words necessary to elicit in our tasks, to be representative of our subjects' lexicons.

The comparisons in chapter seven, when Lex30 was compared with a test designed to elicit a similar number of infrequent vocabulary items, the GapFill task, show similar scoring patterns (table 8.3). This appears to indicate, further, that when we compare Lex30 with a similar test, we elicit similar proportions of infrequent vocabulary items.

Table 8.3 Lex30% and GapFill% task scores from chapter seven.

	Lex30 %	GapFill %
Hiroko	43.8	43.6
Kanako	43.6	51.1
Marie	32.4	39.6
Masami	29.1	33.3
Yukari	29.3	26.9

These results provide very tentative support to the claim that the knowledge elicited by Lex30 is a representative sample of the productive lexicon. However, the results also suggest that we need much more experimentation in order to strengthen the claim that the samples elicited by Lex30 are representative given the limited number of responses (a maximum number of 120) it has the potential to generate.

8.3 Vocabulary knowledge and the aspects of knowledge measured by Lex30.

Results from the experimental chapters (8.2.3) appear to indicate that we might be able to isolate vocabulary knowledge elicited by Lex30 from other aspects of vocabulary knowledge. However, a recent approach to L2 acquisition argues that we cannot separate vocabulary from other aspects of language. For Larsen-Freeman and Cameron (2008), "complexity theory forces us to contend with, not ignore, the dynamism of language and all the messiness it engenders" (p.9) because "from a complex systems perspective, flux is an integral part of any system" (p.152). Larsen-Freeman and Cameron (2008) argue that, "because language is complex, progress cannot be totally accounted for by any one factor or by performance in any one subsystem....[and] [t]he complexity in the language-using system arises from components and subsystems being interdependent and interacting with each other in a variety of different ways" (p148). This 'messiness' suggests that vocabulary knowledge cannot be separated from other aspects of language.

Support for the view that language is 'complex' and 'interdependent' also comes from L2 vocabulary literature that considers vocabulary knowledge to be inseparable from other aspects of L2 knowledge (e.g. Henriksen 1999 (2.4.1) and Read 2004 (2.4.2)). A brief survey of some of the papers reviewed in section 2.2 also reflects this 'complex' and 'interdependent' view of productive vocabulary knowledge. Wesche and Paribakht's (1996) Vocabulary Knowledge Scale assumes that if a subject can put the word into a sentence they know it productively. Similarly, Laufer and Paribakht's (1999) Productive Levels Test assumes that if subjects can complete a word in context, they know it

productively, and Laufer *et al.*'s Computer Adaptive Test of Size and Strength (CATSS) assumes that productive vocabulary is elicited in response to the cues provided in their active recall tasks. Webb's testing (2005 and 2007) also appears to view productive vocabulary knowledge as being multi-faceted, with five tasks that are used to demonstrate knowledge of 'nonsense' productive vocabulary items (for orthography, meaning and form, grammatical functions, syntax, and association). As a final example, Laufer and Nation's Lexical Frequency Profile assumes that if subjects can include words appropriately in composition form they know them productively.

Apparently flying in the face of all of the above studies that consider vocabulary knowledge inseparable from context, a fundamental characteristic of Lex30 is that it accesses productive vocabulary knowledge out of context. Various attempts to measure vocabulary knowledge, such as the above, assume relationships between the different aspects of knowledge that make up the construct of lexical competence. However, the results from chapter seven (see table 7.5) suggest that Lex30 is capable of measuring productive vocabulary knowledge without activating multiple aspects of knowledge. In order to describe how different aspects of knowledge relate to each other, we first need to determine what constitutes the construct of lexical competence, including the multi-dimensional construct of productive vocabulary knowledge, and Lex30 has the potential to do that.

Productive vocabulary is one part of vocabulary knowledge and it is also multi-dimensional. Lex30 allows us to access and measure one of the aspects of vocabulary knowledge (productive vocabulary knowledge), and the results from the experimental chapters suggest that we can isolate vocabulary knowledge elicited by Lex30 from other aspects of vocabulary knowledge.

Two frameworks (Bachman and Palmer 1996, and Read and Chappelle, 2001) reflect the view that aspects of knowledge relate to each other in some way and that, therefore, language knowledge is part of a system (Canale and Swain 1980: 7-12) in which aspects of knowledge relate to one another. Bachman and Palmer (1996) argue against tests like

Lex30, while Read and Chapelle (2001), though largely in agreement with Bachman and Palmer, allow for a 'trait' perspective, giving tests like Lex30 a legitimate role. My argument is that it is only after we are successful in accessing and isolating different aspects of vocabulary knowledge that we might then begin to understand (and describe) how these aspects of knowledge develop and interact and "have a properly worked out theory of what factors contribute to lexical competence" (Meara 1996: 37). This section examines these two frameworks in light of the knowledge gained by using Lex30.

Bachman and Palmer (1996), first, reflect the view that language cannot be isolated into discrete aspects of knowledge. Bachman and Palmer (1996) stress the "need to consider language ability within an interactional framework of language use [and that] (t)raits or attributes of learners cannot be viewed independently of the contexts within which these traits will be manifested" (p.123). They argue that a trait such as vocabulary ability can only be viewed within the context in which the vocabulary is used, and that a subject's vocabulary ability should only be measured within a specific authentic situation.

Accordingly, Bachman and Palmer (1996) reflect the view that language can only be assessed 'as a system'. Yet how knowledge of one aspect of L2 knowledge relates to other aspects of knowledge remains unclear within this view that language can be assessed 'as a system'. In Bachman and Palmer (1996), there is no discussion relating to whether or which aspects of the construct of 'language proficiency' are predictive of others. The framework assumes that an individual with a developed L2 vocabulary knowledge also has developed L2 grammatical skills or strong functional knowledge of the L2 in a particular context. The following points summarize Bachman and Palmer's 'areas of language knowledge' (p. 68):

1. Organizational knowledge (the control or organization of utterances or sentences)
    - Grammatical knowledge: knowledge of vocabulary, syntax, phonology/ graphology

- Textual knowledge: knowledge of cohesion, rhetorical or conversational organization

2. Pragmatic knowledge (how utterances or sentences are related to the communicative goals of the language user and to the features of the language use setting)

    - Functional knowledge – knowledge of ideational, manipulative, heuristic, and imaginative functions of language; or to express/ interpret meaning in terms of our experience of the real world, to affect the world around us, extend our knowledge of the world around us, and to use language to create an imaginary world for humorous or aesthetic purposes

    - Sociolinguistic knowledge – knowledge of dialects/ varieties, registers, natural or idiomatic expressions, cultural references or figures of speech.

One example of how language is considered within this framework can be seen by looking at the idea of language 'usage'. Bachman and Palmer (1996) provide a specific example of how researchers ought to measure usage when they suggest that the "language use we observe and sample in our research [should] correspond(s) to real-life language use" (1996:23), implying that researchers need to correlate authentic communicative samples with any 'language use' produced by subjects. Bachman and Palmer's perspective suggests that a subject's language use would then need to be rated according to the extent that it reflects authentic use.

We therefore need to determine, first, how examiners would rate each of these different components (and how they would isolate each aspect remains unknown), and, second, importantly, the extent to which each of Bachman and Palmer's different components might relate to each other (though perhaps measurements would have to be qualitative). The kind of testing Bachman and Palmer appear to favour suggests that testing is

currently able to evaluate all of these different abilities; however, as discussed earlier (6.3 and 7.3), the fundamental problem with tests is that we have little way of knowing what leads subjects to score in the way that they do, simply because there are so many diverse factors that might influence their results.

Read and Chapelle's (2001) framework, second, in particular their description of 'trait' tests, is worth exploring since they claim that a construct definition of vocabulary ability must be specified independently of context, which appears to suggest that we can isolate and test particular aspects of language knowledge. However, there appears to be some confusion surrounding the exact tests that they claim might be successful independent of context (particularly in light of our experimental findings). Read and Chapelle's (2001) framework is consistent with Bachman and Palmer's (1996) in the sense that they stress the importance of construct definition in terms of "skill elements [that can be] specified as characteristics of the tasks in which language ability is demonstrated" (p. 128). Their framework therefore implies that we should consider ability and use along with the need to define aspects of knowledge. Read and Chapelle favour an 'interactionist' definition of the construct of vocabulary ability that refers to trait, context, and how both traits and contexts might interact by means of an individual's metacognitive strategies.

Read and Chapelle (2001) outline three perspectives toward construct definition depending on whether a test is attributable to trait, context ('behaviourist'), or both trait and context ('interactionist'). They note that most existing vocabulary tests are what they describe as trait tests, which "appear to treat vocabulary as a separate component of language knowledge [and...] can be investigated without reference to the functions of words in grammatical structures, text or discourse" (p. 2). They refer to Laufer and Nation's Productive Levels Test and Lex30 as examples of trait tests because both measure vocabulary without any reference to the context in which words might be used. Their behaviourist perspective assesses knowledge of items in a particular test that reflect a particular context: "[T]he constructs of interest may be micro skills like listening comprehension or writing ability, but they are usually embedded in broader

concepts of communicative proficiency for specified academic, occupational or social purposes" (Read and Chapelle, 2001: 15). In an earlier paper, Chapelle (1998) suggests that the trait perspective is defined as context-independent, whereas a behaviourist perspective cannot be defined without considering underlying aspects of knowledge such as the metacognitive abilities a subject needs to "assess the relevant features of context (e.g. level of formality) and decide which aspects of knowledge (e.g. which words) were needed" (p. 43). The interactionist perspective assesses vocabulary "in relation to a specific context of use" (Read and Chapelle 2001: 19). Read and Chapelle suggest that an interactionist perspective pays greater attention to vocabulary assessment than a behaviourist perspective and offers more design options in creating tests that deal with "vocabulary receptively and/or productively... may be either discrete or embedded, while both selective and comprehensive methods of identifying particular lexical items are possible" (p.15).

Read and Chapelle work through the various dimensions of their framework with reference to eight different vocabulary measures (including the Vocabulary Knowledge Scale, LFP, TOEFL, and others.). They propose that validation studies are needed to support construct definitions (with different kinds of evidence required according to whether these definitions are 'trait,' 'behaviourist' or 'interactionist' in their orientation), score reporting methods (either multiple or global) and the appropriate audience for test presentation. However, their framework does not appear to address validity.

Consistent with Read (2000: 7-11), Read and Chapelle (2001:7) view tasks such as the Vocabulary Knowledge Scale, the LFP, and the Productive Levels Test as trait tests and this is important for two reasons. First, because their view again appears to demonstrate that some authors consider that "a test can present words in quite a large amount of context and still be a discrete measure" (Read 2000:10). Second, they attempt to refer trait principles to construct definition when they claim that "the central construct validity question [for a trait test is] ..does the test measure the underlying characteristics without any influence from the test context?" (2001: 21), while still including the Productive

Levels Test and the LFP as examples of trait tests in their appendix. It appears obvious why Lex30 might fit this description, because there is no consideration given to context but it is less obvious why Read and Chapelle would include the Productive Levels Test and the LFP as examples of trait tests, in view of the context-engaging nature of those texts, though one assumes that this is because they are scored on production of infrequent vocabulary only.

If we measure vocabulary knowledge in the way that these two frameworks (Bachman and Palmer 1996, Read and Chappelle 2001) imply, we cannot tease apart the many factors that might lead to a subject's score. In other words, once we include specific contexts in testing it is not only vocabulary knowledge that influences scoring. The knowledge activated by a vocabulary knowledge test that includes context includes many factors, besides vocabulary knowledge.

In chapter seven, we found that Lex30 scores ranked the same as the Productive Levels Test scores. A possible difference between the two tasks is that, for the Productive Levels Test, consideration might be given to numerous aspects of knowledge (including how a word is written and spelled, in what patterns the word occurs, in what patterns the word must be used, or what types of words can be expected before or after the word, where this word can be used, what word should be used to express this meaning (Fitzpatrick 2007)). Lex30 scores, by contrast, might reflect consideration given to minimal aspects of knowledge (how is a word written and spelled, what word should be used to express the meaning (Fitzpatrick 2007)). The differences between these results appear to demonstrate that we can isolate the vocabulary knowledge elicited by Lex30.

One of the benefits Lex30 has over other tests is that it makes so few assumptions about the developmental relationships between the few aspects of knowledge that it appears to activate, while other tests of productive vocabulary knowledge appear to base their assumptions on the developmental relationships between the aspects of knowledge activated. Given that Lex30 scores are influenced by few aspects of knowledge, we might therefore conclude that Lex30 scores are only likely to be influenced by each

subject's productive vocabulary knowledge. Conversely, any claim that productive vocabulary is elicited in isolation by the other tasks of productive vocabulary reviewed in this thesis appears weak because of the obvious influence from other aspects of knowledge. The experiment with the GapFill task (chapter 7) informs this discussion, as we discussed in section 8.2.3, and further evidence to support the claim that only Lex30 and the GapFill tasks activate only productive vocabulary is shown by the very similar proportions of infrequent vocabulary items elicited by the two tasks.

As seen in the review of Read (2004) (2.4.2), the idea that tests reflect 'language as a system' might confuse the assessment of productive vocabulary, particularly as we are less able to describe it within such a multi-faceted and multi-dimensional framework. While language obviously *is* a system, we first need to determine the extent to which we can separate its discrete aspects and, ultimately, then work out how the different aspects integrate into the whole. In principle then, we ought to begin by determining whether we can separate discrete aspects first. Long before we can make any grandiose claims about being able to measure language ability, as a complex whole, we need to determine the extent to which the separable aspects relate to one another. The encouraging results from chapter seven, discussed above, appear to suggest that we can, at least, isolate the productive vocabulary knowledge elicited by Lex30.

Through an examination of the results from chapter seven, this section has shown that other tasks of productive vocabulary knowledge can activate multiple aspects of knowledge, amongst other things, because they reward considerations given to context. These considerations given to context appear to muddy any claims that test scores are reflective of only productive vocabulary knowledge. The different aspects of L2 knowledge that are activated by other so-called tests of productive vocabulary knowledge (see table 7.5) appear to be based on a different, and perhaps less exact, definition of productive vocabulary ability.

Read and Chapelle (2001) suggest that their examples of 'trait' tests measure vocabulary ability independent of context. Yet the results from the experiment reported in chapter

seven suggest otherwise. The problem with tests such as the Vocabulary Knowledge Scale, the Productive Levels Test, and the Lexical Frequency Profile is that the scores on one of these tests might be attributed to a number of different factors, not just productive vocabulary ability, and are likely to have been informed by contextual knowledge. Tests instead need to be identified that successfully activate the construct that they claim to measure with minimal interference from other aspects of knowledge, and Lex30 appears successful in this regard. The discussion section in chapter seven (7.3) indicated that in comparison with Lex30, other tasks designed to elicit productive vocabulary ability appear to have a greater number of 'activation properties' and to address different aspects of knowledge. In contrast, the evidence from the experimental chapters in this thesis shows that a Lex30 score is only likely to have been influenced by productive vocabulary knowledge.

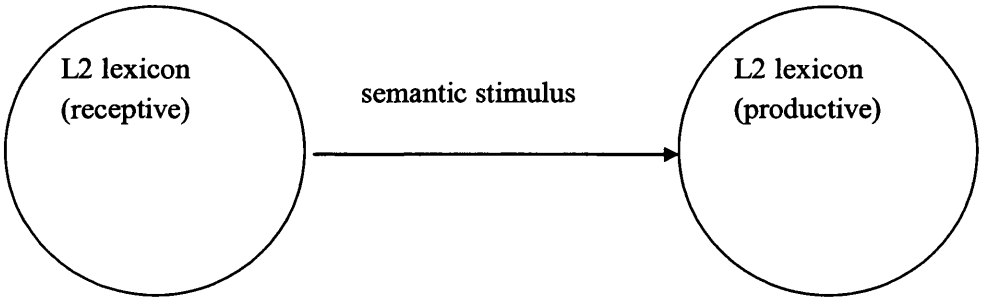
What seems superficially a straightforward construct (i.e. productive vocabulary) actually seems to have several interpretations. When we compare Lex30 to other tests that purport to be testing productive vocabulary knowledge, we see that different tests seem to access slightly different interpretations of the construct. The discussions in chapter seven (7.3), and above, suggest that different performances on the different tests were likely due to the many different activation properties. We see evidence for this claim once we examine the tests we have used in the experimental chapters in light of their different activation properties. Figure 8.1 (adapted from Fitzpatrick 2007: 128) illustrates this by presenting various tests in terms of the way they activate the production of target items. I am using the same model as Fitzpatrick (2007) but have developed it further to include the other tests. Lex30 and the Brainstorm Frequency Profile share the same singular stimulus: an L2 semantic stimulus. Both utilise semantic stimuli: the Brainstorm Frequency Profile task activates items by a sentence and Lex30 activates items by single word stimuli. The GapFill task has two activation properties: an L2 semantic stimulus and an L2 positional stimulus. The Productive Levels Test has three activation properties: L2 semantic stimulus, L2 orthographic stimulus, and an L2

positional stimulus. The LFP has four activation properties: L2 semantic stimulus, L2 orthographic stimulus, L2 positional stimulus, and a composition genre stimulus.

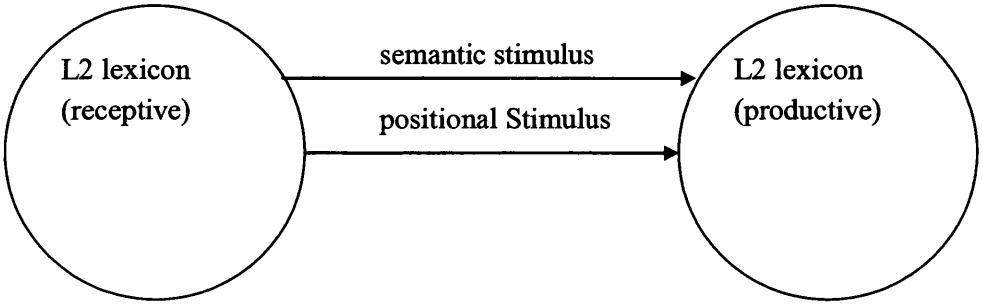
Considering that these five tests activate knowledge in such divergent ways helps us to understand why subjects respond so differently to different tasks. For instance, when we compare Lex30 and GapFill scores in terms of the activation properties we see why the subjects performed differently. Although subjects were not asked to provide grammatically and semantically appropriate responses to the GapFill task, the test data showed that they did.

Figure 8.1 Models of activation for the Lex30, the GapFill task, the Brainstorm Frequency Profile task, the Productive Levels Test, and the LFP task.

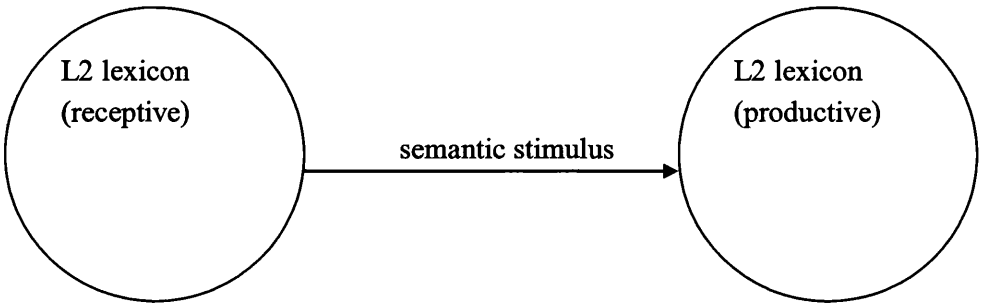
Lex30



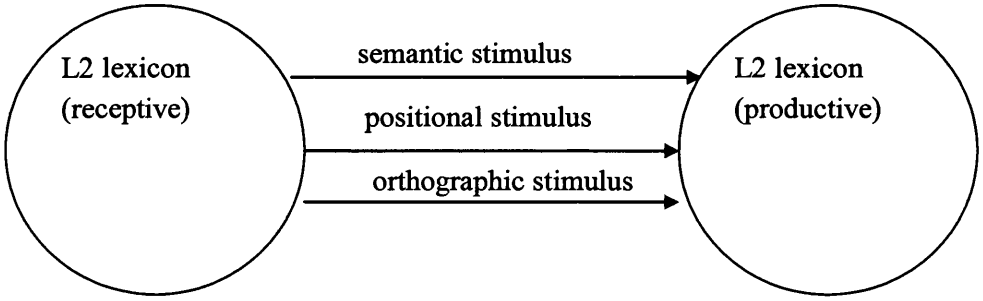
The GapFill task



The Brainstorm Frequency Profile task



The Productive Levels Test



The Lexical Frequency Profile task

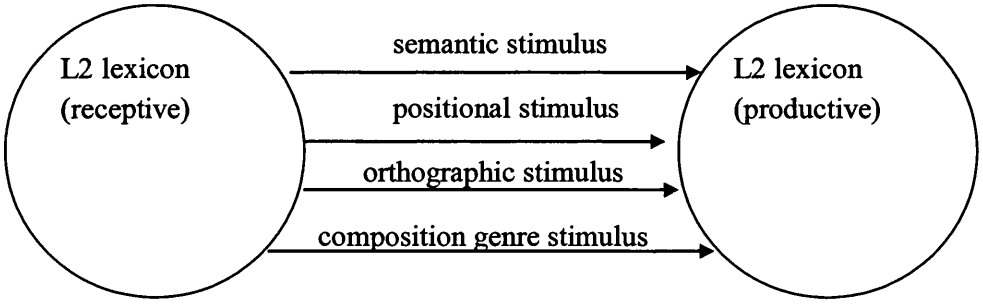


Figure 8.1 represents the modes of activation, or the ways in which knowledge is accessed. The following table examines the tests in terms of Nation's (1990) aspects of knowledge (table 8.2), and represents the aspects of knowledge that are measured by each, the second means, discussed in 7.3, of approaching the question of whether the

tests measure similar aspects of knowledge. Even though each of the five tests claims to measure productive vocabulary, the construct is clearly interpreted in different ways. For instance, both the Productive Levels Test and the LFP include aspects of receptive knowledge. In addition, the Productive Levels Test and the LFP task access a greater number of aspects of knowledge than do the Lex30, GapFill, and Brainstorm Frequency Profile tasks. The Productive Levels Test accesses six different aspects of knowledge, and the LFP accesses seven different aspects of knowledge. Lex30, the GapFill task, and the Brainstorm Frequency Profile only access three aspects of knowledge.

Table 8.4 Aspects of word knowledge (adapted from Fitzpatrick 2007 and Nation 1990) tested by Lex30, the GapFill task (GF), the Productive Levels Test, the LFP, and BFP.

Aspect of Word knowledge (R=receptive, P=productive)			Lex30	GF	PLT	LFP	BFP
Form: Spoken form	R	What does the word sound like?					
	P	How is the word pronounced?					
Form: Written form	R	What does the word look like?					
	P	How is the word written and spelled?	✓	✓	✓	✓	✓
Position: grammatical position	R	In what patterns does the word occur?			✓	✓	
	P	In what patterns must we use the word?			✓	✓	
Position: collocations:	R	What words or types of words can be expected before or after the word?			✓	✓	
	P	What words or types of words must we use with this word?					
Function: frequency	R	How common is the word?					
	P	How often should the word be used?					
Function: appropriateness	R	Where would we expect to meet this word?					
	P	Where can this word be used?			✓	✓	
Meaning: concept	R	What does the word mean?					
	P	What word should be used to express this meaning?	✓	✓	✓	✓	✓
Meaning: associations	R	What other words does this word make us think of?					
	P	What other words would we use instead of this one?	✓	✓		✓	✓

While Nation's (1990) table is important to highlight the differences in what is measured, it is also important to note that the compilation of this table is not so straightforward. For instance, while Lex30 has a tick for productive knowledge in the 'how is the word written or spelled' category, the task demands are weak in this respect because any word is accepted as long as it is spelled approximately. This is the same for the other four tasks but serves to demonstrate the point that the ticks are not straightforward. A tick for 'what word should be used to express this meaning' for both Lex30 and the Productive Levels Test emphasises this point. For Lex30, words are not rejected if the meaning link is unclear, yet for the Productive Levels Test there is only one possible answer for each of its sentence completion tasks. Hence, the test demands for the Productive Levels Test for this category are strong, yet weak for Lex30. Lex30 does not measure receptive knowledge of 'in what patterns we must use the word' whereas the Productive Levels Test does so in its sentence completion tasks. Similarly, productive knowledge of 'where can this word be used' is required by both the Productive Levels Test and the LFP. This analysis shows the complexity involved in measuring productive vocabulary knowledge. In short, this brief analysis indicates that the testing of productive vocabulary knowledge encompasses a range of measures and reflects a range of interpretations.

8.4 Lexical Processing and Lex30: Influence of word frequency.

A Lex30 score is based on a measurement of the number of infrequent (1k+) items produced in response to a list of frequent cue (1k) items. This section examines the connections subjects make between less or more frequent Lex30 cues. A test-taker meeting one of the frequent Lex30 cues responds with associations in their lexicon based on the links they activate. Better known words, likely to be highly frequent items, might have a greater number of connections between them and the connections or links between these items might be stronger than lesser known items with fewer and weaker connections between them within the word web. Repeated activation strengthens the connection and frequent items are more likely to be activated more often. In an early

Lex30 study, Fitzpatrick (2003) found that single word responses to the 100 pilot Lex100 cues failed to 'differentiate [between subjects] in a meaningful way' (p. 121), because the test produced too many frequent responses (in which subjects responded with one word per cue for Lex100). Lex30 aims to "give subjects as much opportunity to produce infrequent responses" (2003: 121) as possible, because subjects are able to respond with more than one response to each cue. First responses to the cues might have strong connections (and activation properties) to other words in the web, while subsequent responses to the same cue might have fewer, weaker, or no other links. Viewing subjects' lexicons in this way enables an exploration of the ways in which subjects access words in their lexicons.

With each generated Lex30 data sample, we might be able to examine the items produced and hypothesize about the nature of a subject's productive lexicon. Considering that the Lex30 cues activate other words in individual lexicons we might see how lexicons, for instance, might be built up of a 'network of associations between [a] word[s] and other words in the language" (Meara 1996: 47). Meara (2006: 625) suggests that this network consists of 'activated' (or productive) and 'unactivated' (or receptive) words, of which both depend on the way they interact with other words in the lexicon. Thus, words within the lexicon are activated once the other words to which they are linked become activated.

Support for this theory about activations and connections that underpin a subject's productive lexicon exists in the data in the experimental chapters. Meara (1983: 30) suggests that high frequency words tend to produce high frequency responses. Data taken from chapter five, comparing Lex30 scores with different sets of cues, and repeated here in table 8.5 below, show that lower frequency cues (JC2k) tend to prompt a smaller number of lower frequency responses from the subjects' lexicons in tests taken at two different test times. However, this is not conclusive because the subjects actually provided fewer responses to the 2k cues, meaning that we would expect these raw scores to be lower as a result. Nevertheless, if we take 1k cues to be frequent and 2k cues to be

infrequent then the results from chapter five suggest that infrequent cues do not elicit more infrequent responses.

Table 8.5 Comparing means and standard deviations of JC1k and JC2k scores (from Chapter 5).

Test time	Task	Mean (sd)
Test time 1	JC1k	23.3 (7.9)
	JC2k	21.7 (5.5)
Test time 2	JC1k	26.5 (7.6)
	JC2k	24.3 (5.9)

More frequent cues appear likely to be more familiar to subjects and therefore have the potential to elicit a greater proportion of infrequent responses. In other words, subjects meeting one of the 1k cues might respond with a greater proportion of infrequent items from their own lexicon and the ability to do this is greater, potentially, than when subjects meet and respond to one of the 2k cues. More frequent cues (from the first thousand frequency band) might have a greater number of connections with other words in the lexicon, and these connections might form a greater number of links than those items elicited from less frequent cues (from the second thousand frequency band) which have fewer connections between them within the word web.

If we examine one subject's Lex30 responses, we see the kinds of responses the lower and higher frequency cues generate. Lex30 data from subject one from chapter five is shown below (see Figures 8.2 and 8.3) in which scoring items are shown in bold and italics. The following section analyses these responses in greater detail, but they are included here to show the kinds of responses elicited by different sets of cues (selected from different frequency (1k and 2k) bands)). In principle, the responses given to the cues tend to suggest that there is a greater proportion of predictable responses to the higher frequency (1k) cues ('brush' ➔ *teeth, socks, paste, cleaning* (*tooth, socks,* and *clean* are amongst the most common responses according to the Edinburgh Associative

Thesaurus (EAT) (http://www.eat.rl.ac.uk/)); 'flag' ➔ *wave, red, white, Japan* (*wave, red,* and *white* are amongst the most common EAT responses); 'head' ➔ *hair, face, cap, large,* etc.) (*face* and *hair* are amongst the most common EAT responses) as compared to the lower frequency (2k) cues (*"*'boundary' ➔ *korea, sea, air* (none of which are amongst the most common EAT responses); "goal' ➔ *finish, happy, attain* (none of which are amongst the most common EAT responses). The scores (table 8.5) show that subjects tend to produce higher scores in response to the higher frequency cues, demonstrating that higher frequency cues prompt a greater proportion of low-frequency responses. If the network development model is operating, we might expect that any newly acquired word will link itself to words already established in the lexicon and that the established words are reasonably likely to be frequent ones. Thus, it is possible that, in the developing L2 lexicon, infrequent (newly acquired) words will be linked to frequent (established) ones. Accordingly, the use of a frequent cue has a higher chance of prompting an infrequent response, while an infrequent cue might generate a frequent word in response.

Figure 8.2 Chapter 5 subject one responses to JC1k – test time one (with bolded and italicised scoring items).

1. away	go, cold, drink, ***bicycle***
2. blow	cold, wind, walk, away
3. brush	***teeth***, ***sock***s, paste, cleaning
4. chance	try, ***shot, tennis***, decide
5. common	friend, knowledge, sense, ***fun***
6. dance	***ballet***, beautiful, ***swan, lake***
7. district	lake, large, quiet, street
8. ever	continue, before, love, unchanged
9. famous	hg, ***talent***, peyonjun, tv
10. flag	wave, red, white, japan
11. get	money, lover, ***bonus***, good
12. head	hair, face, ***cap***, large
13. insect	***spider, dirty, ugly, bug***
14. knee	hit, ***bruise***, sit, ***pad***
15. list	watch, tennis, ***slice***, cut
16. mat	***bath***, front door, wash, dry
17. mountain	climb, nozato, mountain, ***ski***
18. oil	mother, king, rule, ***elder***
19. pattern	check, ***uniform***, life, ***dull***
20. policeman	***bike, scary***, white, ***panda***
21. public	***library***, free, ***useful, convenient***
22. religion	christian, kobe, college, islam
23. secret	game, cold, hard, back
24. shirt	***skirt***, summer, ***vacation***, hair
25. sorry	bad, ***excuse, angry, plea***
26. smell	***stink, rotten***, socks, ***disgust***ing
27. spirit	quickly, german, ***pen, pencil***
28. surprise	***shame***, study, effort, ***god***
29. telephone	mother, handy, call, friend
30. tool	april, let, ***stupid***, dull

Figure 8.3 Chapter 5 subject one responses to JC2k – test time one (with bolded italicised scoring items).

1. affect	movie, cat, move, love
2. area	***territory***, cat, radio wave,
3. balance	ball, ***seesaw***, un***stable***,
4. boundary	korea, sea, air, fly
5. cement	hard, white, ***slick, skate***
6. comment	***severe, tennis, advice***, tv
7. connect	***consent***, tv, breaker, ***defective***
8. court	tennis, judge, ***clay***, run
9. degree	***temperature***, study, chemistry, hard
10. dismiss	***restrict***ion, no, homeless, park
11. energy	power, plant, nature, work
12. extreme	***tired***, science, ***busy***, changing
13. flow	water, ***bath, pool, electricity***
14. goal	finish, happy, ***attain***, good
15. hook	***cherish***ed, hand, arm, lose
16. index	***library***, card, finger, note
17. just	late, long, right, ***online***
18. load	heavy, ***freight***, car, ***track***
19. memory	lose, study, bad, full
20. oblige	present, job, school, ***uniform***
21. pain	***bruise***, cut, ***mental, sad***
22. point	finger, ***blackboard***, win, game
23. profession	doctor, teacher, ***nurse***, scientist
24. reaction	fast, run, friend, ***cool***
25. research	chemistry, hard, ***internet***, bad
26. sale	run, ***supermarket***, monday
27. ship	sea, wave, titanic, island
28. sport	***soccer, volley***, news, ***skate***
29. suit	arbeit, ***interview, ceremony***, black
30. tight	***rope, pants, schedule, choke***

When scoring the tests, the subjects' responses were lemmatised according to Bauer and Nation's second and third level affixation list. Hence, '*handy*' (the second response to the 29th Lex30 1k cue *telephone*) was lemmatised to *hand* because the adjective is the derivative of that noun, adding the level 3 affix '-y'. No proper nouns were counted as scoring items, which is why, for example, the responses to the 22nd Lex30 1k cue *religion* (*Christian, Kobe, College, and Islam*) were not included as scoring items. Repeated words do not score repeatedly, hence the response *lake* scores only once (although it was given twice, to the 6th and 7th Lex30 1k cues). Finally, items that are not English are not treated as scoring items, as was the case with the responses to the 9th Lex30 1k cue *famous* (i.e. *hg, peyonjun*). A detailed explanation of the scoring procedure is provided in section 3.2.3.

As table 8.6 shows, the scores for subject one are lower (both as raw and percentage scores) for cues selected from the second thousand frequency band than for cues selected from the first thousand frequency band. The difference between scores for subject one is small, but the mean differences (table 8.5) reflect a consistent difference between scores (and the standard deviations are suggestive of less variability between subjects with the JC2k cues).

Table 8.6 Comparison of 1k and 2k scores from chapter 5 – subject one.

	JACET8000 raw		JACET8000 percentage	
Cues	1k	2k	1k	2k
Score	43	42	36%	35%

This section discussed the connections subjects make between more frequent and less frequent Lex30 cues. Better-known cues might inherently activate a greater number of connections than lesser-known cues that generate fewer connections.

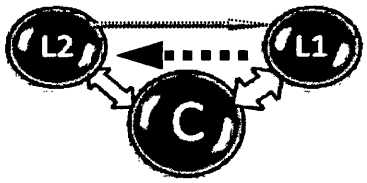
8.5 Lex30 and a model of the bilingual lexicon.

This section discusses findings from the experimental chapters in the context of a model of the bilingual lexicon (in which the term 'bilingual' "refer(s) to anyone who uses a

second language at a relatively high level of proficiency" (Kroll *et al* 2002: 138)), and aims to determine whether the model might explain the ways in which L1 words, L2 words and concepts might interact when the test-taker responds to Lex30. This section also examines how a network of activation is operating when subjects respond to Lex30. The section is divided into two parts. The first part hypothesises a bilingual model in terms of Lex30 response behaviour. The second part examines the nature of the connections that subjects make in response to the Lex30 cues. This second part examines responses to Lex30 in terms of whether they tend to be 'semantic' or 'lexical' in order to indicate the kinds of networks that might be activated in response to the Lex30 cues.

A bilingual model might help us to understand the way in which items are integrated into the lexicon. A Hierarchical bilingual model assumes that bilinguals organize their languages into one general conceptual level, shared by the two (the L1 and L2) languages, and a lexical level that is particular to each language. The models are hierarchical in the sense that they distinguish between the discrete different levels: a conceptual level and a language specific mental lexicon. The conceptual level is represented by the circles in the diagrams (as in figure 8.4 below) labelled 'concepts,' while the lexical level is represented by the circles labelled 'L1' and 'L2', for the first and second languages. The shared conceptual level system contains general abstract information that is language free, while the lexical level represents the bilingual's two languages in separate lexicons with information specific to each language. The following describes the hierarchical model in order to hypothesise about Lex30 response behaviour.

Figure 8.4 Kroll and Stewart's (1994) Revised Hierarchical Model.



Kroll and Stewart's (1994) Revised Hierarchical Model (RHM) (figure 8.4) proposes a series of connections of various strengths between words and concepts in the L1 and L2. With the RHM model, there is a shared conceptual store, from which language specific lexicons are accessed via translation as well as conceptual mediation. The model therefore accounts for the possibility that L2 learners continue to rely on their L1 irrespective of their L2 proficiency. The RHM model also accounts for developmental shifts in the sense that, as individuals become more proficient in their L2, they tend to rely less on mediation as they become more able to access concepts directly. A central aspect of the RHM is its asymmetrical nature, assuming that, because of the way second languages are learned, all L2 words connect to L1 words, but not all L1 words necessarily connect to L2 words.

Therefore, for the RHM, L2 words are often thought to be learned initially by associating them with their L1 translation (such as learning the L2 Japanese て (te) from the L1 English 'hand'). The connections from L1 to L2 words are not thought to be particularly strong since learners tend not to use their L2 in this way (i.e. there is absolutely no reciprocal need to access the L2 Japanese 'て' (te) in order to find the L1 English word 'hand', for a native English speaker learning Japanese). Therefore, there is a stronger lexical link from L2 to L1 (explaining why backwards translation (L2 to L1) is faster than forwards translation (L1 to L2)), and a stronger conceptual link between the conceptual store and the L1 lexicon.

Figure 8.5 Kroll and Stewart's (1994) RHM and Lex30.

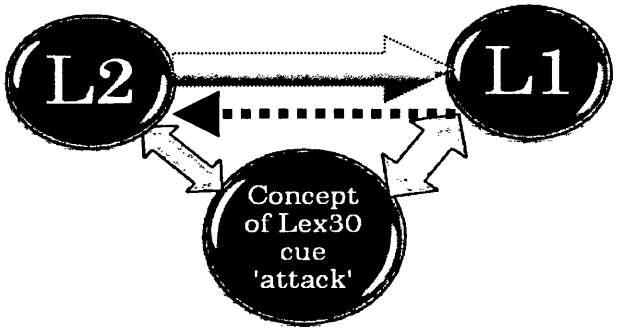


Table 8.7 presents a brief analysis of individual responses (taken from the replication study reported in chapter three) to the first Lex30 cue 'attack' in terms of the RHM (figure 8.5). The three subjects range in terms of their L2 English proficiency as follows: Kenji is a post-elementary English learner, Ryohei is a pre-intermediate English learner, and Yusuke is a post-intermediate English learner.

Table 8.7 Three different subject responses (taken from the replication study reported in chapter 3).

Subject	Cue	Responses
Kenji	attack	*offence, block, damage, volleyball*
Ryohei	attack	*army, problem, heart, war*
Yusuke	attack	*tiger, terrorism, cat, my leg*

We might hypothesize that Kenji's responses to the Lex30 cue 'attack' seem to have a dominant volleyball theme. There is a comic about volleyball in Japan, called 'Attack', so the concept of 'attack', and the direct L1 translation equivalent of 'attack' might not have been accessed here. Instead, we appear to have evidence of L2-L2 intralingual links. In contrast, Ryohei might have translated some of his responses from the L1 (*army, problem, war*), while *heart* might have been accessed directly from the conceptual store without the need for translation (given that the RHM suggests that as individuals become more proficient in their L2, they rely less on translation). Accordingly, the responses from Yusuke, the subject with the most advanced English proficiency of the three, might be examples of words that have been accessed directly from the conceptual store or through the L1.

Bilingual models might also help demonstrate that subjects sometimes respond with false cognates to Lex30, as discussed in section 3.3 and section 4.4. This analysis of bilingual models appears to suggest that responses might best be explained as generating particular concepts. This includes example responses such as *ufo* in response to *pot* (*ufo*

being the name of a popular brand of noodle in Japan that comes in a bowl-shaped 'pot'), *consent* (a loan word for *plug*) in response to *connect*, or *arbeit* (a German L2 loan word for temporary worker) in response to *suit* (table 8.8). Such responses appear to highlight that the subjects' concepts of such cues reflects L1 rather than L2 knowledge. This is a complex issue that the Revised Hierarchical Model helps to illustrate.

Table 8.8 Examples of potential false cognates.

Lex30 cue	Potential L1 response
cloth	*wear*
pot	*ufo*
connect	*consent*
suit	*arbeit*

The Revised Hierarchical Model might help us understand the potential processing routes for subjects' responses to Lex30 and suggests that there is a degree of interaction between L1 and L2 words and concepts, possibly relating to L2 development. This resonates with Kroll and Stewart's (1994) suggestion that the less proficient subjects are, the more they appear to access L2 words by translating from their L1 equivalents. Having looked at some of the particular connections that subjects provided, a picture begins to emerge of how learners are accessing and producing scoring words in response to Lex30. It is also apparent that some responses to Lex30 might be examples of false cognates.

Next, I aim to examine whether bilingual models explain whether a network of activation is operating. If we examine Lex30 subject response data, we begin to see the kinds of networks that the cues might be activating. The hierarchical model reflects the view that there are separate levels of representations for words and their meanings (Kroll and Stewart 1994). A central issue that the RHM model might help to explain is how words are linked to their meanings across the languages used by bilinguals. According to the RHM, words from each language are interlinked at the lexical level, but the link

from L2 to L1 is stronger than the link from L1 to L2 (seen in the example of the accessing of the L1 Japanese ' て' from the L1 English 'hand'). This linkage reflects the fact that, during L2 learning, translations are made from the L2 to L1 in order to access meaning. The link between L1 words and their meanings is strong, and the link between L2 words and their meanings develops with an increase in L2 fluency. The RHM helps to hypothesize how bilinguals organize their languages and depicts the structure of lexical representation as proficiency increases. If, therefore, we examine subject responses to Lex30, we might see how the nature of the relation between words and their meanings changes as a function of fluency in each language. As language proficiency increases the RHM implies that the connection between an L2 word and its meaning becomes more direct. The RHM suggests that, at an early stage of acquisition, lexical links between L1 and L2 words are stronger than conceptual links between the concept and its corresponding L2 word. Kroll and Stewart's (1994) theory, that early bilinguals appear to use the word association link, while more proficient bilinguals show greater use of concept mediation, suggests that Lex30 responses might indeed be processed in this way, which I now examine.

Thus, the following examines three subjects' Lex30 results (Kenji, a post-elementary L2 English learner, Ryohei, a pre-intermediate English learner, and Yusuke, a post intermediate English learner). The analysis (see figure 8.6) examines each subject's responses in terms of the proportion of lexical and conceptual links that they provide in response to the (first five) Lex30 cues, although I should add that these are hypothesised and therefore do not represent accurate or confident categorizations. If subjects produce what appears to be a collocation it is possible that they have not mediated the cue or response through the L1. Hence figure 8.6 indicates the proportion of L1 mediated (not collocational) or not L1 mediated (collocational) links that the subjects appear to have made. Given that collocation links are lexical, and everything else is 'conceptual' figure 8.6 suggests that the more advanced learners make more collocational links.

Figure 8.6 An analysis of three subjects' responses to the first 5 Lex30 cues.

Potential collocation links indicated by a surrounding box ( )	
Kenji (a post-elementary learner)	
1. attack	*offence, block, damage, volleyball*
2. board	*flat, white, black, snow*
3. close	*door, store, time, window*
4. cloth	*square, thin, stew*
5. dig	*hole, shovel*
Ryohei (pre-intermediate learner)	
1. attack	*army, problem, heart, war*
2. board	*boat, plane, notice, blackboard*
3. close	*window, door, shut, near*
4. cloth	*table, hanger, lay, wear*
5. dig	*hole, shovel, tunnel, potato*
Yusuke (post-intermediate learner)	
1. attack	*tiger, terrorism, cat, my leg*
2. board	*ship, plane, trip, sea*
3. close	*friend, cleaning shop, my home, convenient*
4. cloth	*warm, coat, t-shirt, pants*
5. dig	*deep, mad, tired, grass*

This brief survey suggests that different kinds of networks are being activated by the Lex30 cues and reveals that a developmental pattern does appear to be in operation. The

analysis suggests that the more proficient the L2 subject is in their L2 development, the more they appear to make use of the collocational link.

8.6 The construct of productive vocabulary and Lex30.

We saw in the revised version of Nation's (1990) table (in table 8.9) that Lex30 might mainly access the written and spoken knowledge of responses, while also, potentially, accessing knowledge of the form and meaning. I outlined a number of potential questions at the end of the literature review (section 2.5) and sought to answer those in the experimental chapters (3 to 7) to determine precisely what it is that Lex30 measures.

A brief summary of the findings from the experimental chapters is presented here and is used to identify and examine exactly what it is that Lex30 measures. Based on the literature review (chapter 2) and experiment chapters (3 to 7), I might make four broad claims about what Lex30 measures.

- Lex30 appears broadly successful at accessing subjects' productive vocabulary knowledge regardless of the mode (written or spoken) in which they respond to the task. The concern about whether subjects with lower levels of orthography might be at a disadvantage when confronted with the written version of the Lex30 test appears unfounded given that there were only negligible differences between the spoken and written forms of subjects' Lex30 test results (see section 4.3).
- Lex30 cues selected according to the same criteria as Meara and Fitzpatrick (2000), and from the same frequency band, consistently elicit samples of subjects' productive vocabulary ability (see section 5.3), as seen by the similar scores from the two versions of Lex30 (with different sets of cues taken from the same 1k frequency band). This analysis certainly provided sufficient evidence to claim a degree of reliability for Lex30, and suggests that, even with alternative cues, Lex30 successfully taps subjects' productive vocabulary abilities.

- Lex30 appears to be valid since test scores improved following a period of six weeks of English instruction. Regardless of the (1k or 2k) frequency band from which the cues were selected there was a significant pattern of difference between subjects' scores at each of the two tested times (see section 5.3), indicating that the subjects produced a greater number of infrequent productive vocabulary items at the second test time.
- Lex30 differs from other tests in the sense that it appears to access only minimal aspects of productive vocabulary knowledge and does not access multiple aspects of lexical knowledge (see sections 6.3 and 7.3). When I designed the GapFill task, a task with similar activation properties to Lex30, I found that the similarity between Lex30 and GapFill task scores appeared to be due to the small number of activation properties and the same aspects of knowledge that are accessed. The results from chapter seven, therefore, offer cautious but encouraging support for the construct validity of Lex30 as a test of 'productive vocabulary'. Other support for construct validity comes, albeit cautiously, from the changes in test performance over the six week period reported in chapter five, and the similarities between the scores on the spoken and written versions of Lex30 reported in chapter four.

The majority of tests of productive vocabulary appear to be accessing multiple aspects of knowledge that may either help (if, for instance, the particular test is reflective of subjects' learning paths) or hinder subjects' scores (if the particular context is unreflective of subjects' knowledge). The danger remains that some tasks appear to measure "aspects of the task that are extraneous to the focal construct mak[ing] the test irrelevantly more difficult for some individuals or groups" (Messick, 1989: 34). Lex30 appears to make no such extraneous demands on its subjects, so it appears unlikely that the limited aspects of knowledge that are elicited by Lex30 will make the test 'irrelevantly... difficult' for particular groups or individuals. Thus, Lex30 offers much wider applicability.

The analyses reported in chapters six and seven support the suggestion (Fitzpatrick 2007) that there might be some danger in grouping tasks under the umbrella heading of 'tests of productive vocabulary knowledge'. Such grouping appears to stem from the fact that some authors (Bachman and Palmer 1996, Read and Chapelle 2004) appear to view productive vocabulary as only accessible in context and via multiple aspects of knowledge. This view is reflected by the many tests designed to elicit productive vocabulary knowledge (see table 2.10). Table 8.4 shows that the number of different aspects of vocabulary knowledge that Lex30 accesses is smaller when compared to other tests, but this is by no means a bad thing. In section 8.3, we saw that other tests of productive vocabulary knowledge reflect Read's view that 'a test can present words in quite a large amount of context and still be a discrete measure' (2000:10). However, as discussed in section 8.3, the danger of testing in this way is that such tests might fail to capture knowledge of items that subjects may know. This might be due to the particular items that examiners choose to test, or to the particular context that is given being unreflective of a subject's learning path.

If we look at the way Lex30 tests, we can examine which aspects of knowledge it measures. Subjects respond to Lex30 by providing any word that they can think of in response to the highly frequent cue words. Any response is accepted as long as it is spelt approximately and not rejected if it is deemed inappropriate. There is usually some sort of meaning link, suggesting that some semantic or conceptual access takes place. The number of infrequent items that subjects produce in response to Lex30 is taken to relate to the number of infrequent items that the subject is assumed to know. Subjects' Lex30 scores are based only on their production of infrequent items, so a high Lex30 score is one in which a subject has produced many infrequent items, while a low Lex30 score is one in which a subject has produced few infrequent items. Lex30 therefore measures subjects' ability to produce infrequent items, and a response to Lex30 implies a minimum level of productive vocabulary knowledge of form, and probably some semantic or conceptual connection, too.

If we return to Nation's table, adapted by Fitzpatrick (2007: 129), we might now re-evaluate the aspects of knowledge that Lex30 activates (table 8.9). Based on the results in chapter four (see 4.3), we can add productive knowledge of how a word is pronounced. This suggested change is included under the 'Clenton' heading. We saw that better known, or more frequent, cues appear to elicit higher Lex30 scores and might have a greater number of connections or links between items within subjects' word webs. In terms of bilingual models and Lex30 (8.5), we see that there is a degree of interaction between L1 and L2 words and concepts, once subjects respond to the Lex30 cues. The examination of bilingual models also appears to account for why subjects appear to produce false cognates (such as *ufo*, or *wear*) and that networks are operating when subjects respond to the Lex30 cues.

Table 8.9 Aspects of word knowledge (adapted from Fitzpatrick 2007 and Nation 1990) tested by Lex30.

Aspect of Word knowledge (R=receptive, P=productive)			Fitzpatrick	Clenton
Form: Spoken form	R	What does the word sound like?		
	P	How is the word pronounced?		✓
Form: Written form	R	What does the word look like?		
	P	How is the word written and spelled?	✓	✓
Position: grammatical position	R	In what patterns does the word occur?		
	P	In what patterns must we use the word?		
Position: collocations:	R	What words or types of words can be expected before or after the word?		
	P	What words or types of words must we use with this word?		
Function: frequency	R	How common is the word?		
	P	How often should the word be used?		
Function: appropriateness	R	Where would we expect to meet this word?		
	P	Where can this word be used?		
Meaning: concept	R	What does the word mean?		
	P	What word should be used to express this meaning?	✓	✓
Meaning: associations	R	What other words does this word make us think of?		
	P	What other words would we use instead of this one?	✓	✓

We might now make five broad conclusions that relate to what it is that Lex30 measures:

- Lex30 appears only to activate minimal aspects of vocabulary knowledge when compared to other tests of productive vocabulary knowledge (see section 2.2) because it does not make assumptions about the learning backgrounds of test takers (i.e., knowledge of infrequent items is taken to mean any item produced beyond the

1k+ level). Other tests appear to test in this way, but the items that are the focus of the test are expected to be learned in a particular order, such as in the Productive Levels Test, or CATSS test. Beyond the 1k level, there is no such assumption in the testing with Lex30. The items subjects provide might only tell us about their threshold knowledge.

- Lex30 appears to tap subjects' knowledge of infrequent vocabulary items regardless of the (written or spoken) response format. The results from the comparison between written and spoken response formats of Lex30 (in chapter 4) show that there was only negligible difference between subjects' scores. This implies that Lex30 defines productive vocabulary as being able to produce an item in response to either the spoken or the written format of the task. In other words, if a subject can respond to the spoken format of Lex30, they can also respond to the written format of Lex30.
- Cautious support for the construct validity of Lex30 as a measurement of 'productive vocabulary' comes from the similarity between Lex30 and GapFill task scores, appearing due to the two tasks' small number of activation properties, as well as the accessing of the same aspects of knowledge reported in chapter seven.
- Cautious support for the construct validity of Lex30 as a measurement of 'productive vocabulary' comes from changes in Lex30 test performance over the six-week period, as reported in chapter five.
- Lex30 appears to activate knowledge of a web of words and we see this at work through two hypotheses. The first hypothesis relates to the storage and access of items elicited by Lex30 (see section 8.4). Cues that are more frequent appear to activate knowledge of infrequent items, while less frequent cues appear to activate knowledge of more frequent items. The second hypothesis relates to the exploration of how bilingual models (see section 8.5) relate to Lex30 response behaviour and suggests that: (1) responses might be explained by the relationship between each subject's L1 and L2 words and their related concepts; and, (2) L2 proficiency might

relate to the proportion of what appear to be either the direct or L1-mediated links that subjects provide.

8.7 Conclusion.

Chapter eight has attempted to draw together the different strands of the thesis and discussed these in five broad sections. The first section addressed three issues related to the knowledge that Lex30 aims to elicit. The first of these related to the percentage and the raw scoring systems, suggesting that while the percentage scoring system assesses each subject's ability to produce infrequent vocabulary items spontaneously, the raw scoring system instead assesses more than one aspect of language competence, including motivation, fluency, and others. I then discussed the use of particular frequency lists that might or might not reflect particular groups of subjects' learning paths. This recognition of such differences is important, and a strength of Lex30, in that it is adaptable because it can use different frequency lists depending on the learners (such as using Jacet8000 for Japanese learners of the L2 (English) or Nation (1984) for L2 learners resident in New Zealand). The third part of this first section examined how effective Lex30 is at sampling productive vocabulary and whether the items that are elicited are genuinely representative sample of the subjects' lexicons.

The second section dealt with three broad areas that relate to Lex30 as a test of productive vocabulary. The first discussed issues related to the consideration of vocabulary as part of general language knowledge. The second related to wider vocabulary knowledge in relation to the aspects of knowledge measured by Lex30 and argued that, only after we are able to assess productive vocabulary knowledge with minimal influence from other aspects of knowledge, we might then be in a position to start addressing how different aspects of knowledge relate to one another. The third considered Lex30 compared to other tests of productive vocabulary and the various interpretations of the construct of vocabulary knowledge. This discussion is intrinsically related to the experimental studies (chapters 3 to 7) in the sense that other measures of

productive vocabulary knowledge might actually have 'different activation properties' (Fitzpatrick 2007: 127) and access different aspects of knowledge.

In the third section, I looked at lexical processing and Lex30. The section began by comparing different sets of cues selected from different frequency bands, hypothesizing that more frequent cues appear to elicit more infrequent responses than less frequent cues.

In the fourth section, I hypothesised about Lex30 response behaviour in light of the Revised Hierarchical Model (RHM). The possibility emerged that the RHM might account for how less proficient learners might be more likely to rely on translating L1 equivalents than more proficient subjects (who might access their conceptual store without accessing knowledge of L1 items). A developmental pattern might also occur in the reflection of the proportion of responses making use of the lexical link in the sense that the more proficient the L2 subject appears to be, the less they appear to make use of the lexical link.

The fifth and final section examined the construct of productive vocabulary and Lex30. Lex30 measures productive vocabulary regardless of whether it is administered in the written or spoken mode. A response to Lex30 implies a minimum level of productive vocabulary knowledge of form. We can analyse subjects' responses, revealing that infrequent cues generate frequent responses, while, conversely, frequent cues generate infrequent responses. An analysis of our subjects' responses might also reveal other information about their L2 proficiency. Subjects who make a greater proportion of, what appear to be, semantic connections to the cues might be more proficient than those subjects who make a greater proportion of, what appear to be, lexical connections.

This conclusion draws together the five sections of chapter eight and shows that Lex30 has the potential to be an important testing tool for measuring productive vocabulary knowledge. We now know that the percentage scoring system measures productive vocabulary knowledge discretely, without the many interfering factors that influence the

raw scoring system. We also know that, depending on the learners that we aim to test, Lex30 is adaptable because we can use different frequency lists that reflect the learning backgrounds of our subjects. Thus, Lex30 also appears to offer a first step in providing the means to access, in isolation, one of the many aspects of language knowledge, productive vocabulary knowledge. In short, Lex30 offers the potential to hypothesise about our subjects' L2 proficiency, not only in terms of the proportion of infrequent items but also in terms of the number of L2 collocations that they provide.

Chapter 9 Conclusion.

I began the thesis by presenting a review of a number of different tests, each claiming to measure productive vocabulary knowledge. The review of these tests, in section 2.2, suggested that, as well as accessing productive vocabulary, many tasks appear to activate multiple aspects of vocabulary knowledge. Lex30, however, appears to access productive vocabulary without activating such additional aspects of vocabulary knowledge. I outlined a number of potential questions at the end of the literature review (section 2.5) and sought to answer these in the experimental chapters that followed (3 to 7) to attempt to determine precisely what it is that Lex30 measures. We saw in the revised version of Nation's 'what is involved in knowing a word' (tables 8.4 and 8.9) that Lex30 might only access the written and spoken knowledge of a particular word whilst also, potentially, accessing knowledge of the form, and meaning.

Following on from the experimental chapters (3 to 7) and the discussion chapter (8) I am now able to make five broad claims relating to Lex30 as a measure of productive vocabulary.

First, Lex30 measures threshold productive vocabulary knowledge (8.3). I criticised existing tests of productive vocabulary because they implicitly assume developmental links between the different aspects of knowledge that constitute lexical competence. Such tests appear to assume that individuals with strong vocabulary knowledge are also likely to have developed abilities in other aspects of L2 knowledge, such as strong grammatical knowledge (Cummins 2000: 123) .While this might not be such a controversial claim, the assumption is potentially misleading because we cannot know which aspects of knowledge influence the scoring on such tests. Lex30 measures minimal aspects of productive vocabulary knowledge without influence from multiple aspects of language knowledge. With Lex30, we might be activating and assessing productive vocabulary knowledge, which is the focus of this thesis.

We might now begin to consider how productive vocabulary knowledge and other aspects of knowledge interrelate, although that is far beyond the scope of this thesis. Meara argues that, because "we don't have a properly worked out theory of what factors contribute to lexical competence" (1996: 37), "can only really develop models of lexical competence [once] [...] [we] have a complete model of semantics and a complete specification of the syntactic and associational behaviour of all the words in a speaker's lexicon" (1996: 50-51). Once we are able to identify each discrete, individual aspect that constitutes lexical competence, then we might determine how, or indeed whether at all these particular individual aspects relate to each other. Lex30 appears to be a useful starting place in terms of attempts to isolate particular aspects of lexical knowledge.

Second, Lex30 does not make assumptions about the content of individuals' lexicons based on the order in which items are learned (see section 8.3). Tests such as the Productive Levels Test and the CATSS test base their testing on the assumption that words are learned in a particular order (L2 learners learn the first thousand words, then the second thousand words, and so on sequentially). Lex30 does not assume any such developmental progression and bases its scoring on knowledge of any word from outside the 1k frequency band. With Lex30, we appear able to identify only the words of which subjects have some degree of knowledge. This might only be threshold knowledge, but with each generated Lex30 sample, we might then be able to examine the items that subjects produce and hypothesize about the nature of their productive lexicons.

Third, Lex30 bases its testing on the assumption that a network structure exists in which words are usually interlinked by semantic association (8.4). By testing with Lex30 and with more frequent (1k) cues that are likely to have many links (compared to less frequent (2k) cues), we might be able to make claims about subjects' productive lexicons in terms of the frequency profile of their entire productive lexicon. This hypothesis of a network structure suggests that more frequent (1k) cues might generate a greater number of infrequent items than less frequent (2k) cues (which appear to generate a smaller number of infrequent items).

Fourth, once we examine Lex30 data, in light of bilingual lexicon literature, we begin to see that subjects might be making lexical or semantic connections to the cues. The proportion of what appear to be lexical or semantic links the subjects make might tell us something about their L2 proficiency. Hence, when subjects make what appear to be a greater number of semantic connections they appear to be more proficient in the L2 than when they make what appear to be lexical connections. In addition, samples from Lex30 subject data suggest an explanation for what appear to be false cognates in response to the task (8.5) (e.g. *pot* ➔ *ufo*).

Fifth, we now have additional support for the construct validity of Lex30 as a measurement of 'productive vocabulary' following Meara and Fitzpatrick (2000) and Fitzpatrick (2007). This support stems from the four sets of results that reflect: i.) similar test scores on parallel forms of the task; ii) changes in test performance over a six-week test period; iii) similar scoring on the spoken and response versions of Lex30; and, iv) similar scoring on a similar task (the GapFill task).

There are still issues that need addressing relating to the accuracy and scope of Lex30 as a test of productive vocabulary. Accordingly, I have identified three main objectives for future research, to:

- improve the cues so that subjects might find it easier to provide a greater number of infrequent responses to a particular set of Lex30 cues. Certain cues elicited lamentably few responses (such as 'substance'), while other cues elicited many highly frequent responses (such as *furniture,* which tended to elicit lexical sets (e.g. *chair, table, desk,* etc.)). One design feature of the GapFill task (reported in chapter 7) was the attempt to avoid lexical sets. The fact that the GapFill% task elicited a greater proportion of infrequent items than Lex30% might have been due to the lack of lexical sets elicited by the GapFill task. Comparing sample sets of cues with different groups of subjects might provide us with a more effective set of Lex30 cues. While the improvement of the Lex30 cues is an important aim, it is not within the scope of this thesis since I have mainly been concerned with

whether Lex30 assesses the construct of productive vocabulary knowledge rather than the extent to which a revised Lex30 test might elicit a greater proportion of infrequent vocabulary items.

- analyse responses (1) in terms of what we can glean from the test. For the moment, we are still unable to say a great deal about what subjects might know about the words that they produce. Section 8.2.3 discussed whether Lex30 might only elicit threshold knowledge. A significant number of the subjects tested within the individual studies reported here provided potential false cognates to the Lex30 cues (such as *ufo* for the cue '*pot*' (*ufo* being the name of a popular brand of pot noodles in Japan)). The discussion relating to Lex30 and the bilingual lexicon (8.5) suggests that we can begin to see the kinds of words that are activated by Lex30 in terms of subjects' concepts. It would be a useful idea to conduct post-task interviews to discover exactly what subjects do know about the words with which they have responded.

- analyse responses (2). The analysis of Lex30 response data in section 8.5 suggests that the proportion of collocational connections that subjects appear to have made might tell us something about their relative L2 proficiency. In other words, subjects who mainly make what appear to be lexical connections in response to the Lex30 cues might be less proficient than those subjects who make what appear to be a higher proportion of semantic connections.

Clearly more work is necessary within this area, as the three suggestions above for future research imply. In the meantime, though, the results from the studies and the related discussions in this thesis suggest that Lex30 provides us with a helpful means to understand the construct of productive vocabulary:

- Lex30 elicits productive vocabulary regardless of the (spoken or written) mode.
- Parallel forms of the test (with different cues) elicit similar scores.
- More frequent (1k) cues elicit a greater proportion of infrequent responses than less frequent (2k) cues, with the latter appearing to elicit a higher proportion of more frequent responses.
- Similar test results on a similar task (the GapFill task) provide cautious support for the construct validity of Lex30 as a test of productive vocabulary knowledge.
- Lex30 appears to activate and elicit productive vocabulary knowledge without the influence of other aspects of language knowledge.
- Lex30 has revealed that productive vocabulary is itself inherently multidimensional.

Appendix 1 Sample data: completed Lex30 test Subject: 1 (Chapter 3)

1.	attack	*defence, offence, gun*
2.	board	*score, snow, bill*
3.	close	*fire, up, long*
4.	cloth	*ware, belt, saw*
5.	dig	*mine, donkey*
6.	dirty	*clean*
7.	disease	*increase*
8.	experience	*knowledge*
9.	fruit	*diet, meat, vegetable*
10.	furniture	
11.	habit	*manner*
12.	hold	*up, grab*
13.	hope	*desire*
14.	kick	*back, punch, chop*
15.	map	*way*
16.	obey	*speech*
17.	pot	*saucer*
18.	potato	
19.	real	*virtual, fantasy*
20.	rest	*foot*
21.	rice	*wheat, flour, ball*
22.	science	*alchemy, fairy*
23.	seat	*box*
24.	spell	*curse, magic, wizard*
25.	substance	*enzyme, material*
26.	stupid	*wise, dull, clever*
27.	television	
28.	*tooth*	*mouth, ear, eye*
29.	trade	*building, terrorism, world*
30.	window	*computer, apple*

Appendix 2 Lemmatisation criteria (Meara and Fitzpatrick 2000: 29-30)

Words were lemmatised according to the criteria for level 2 and 3 affixes described in Bauer and Nation (1993). Words with affixes included in the lists below were treated as instances of their base lemmas, and scored accordingly. Words with affixes that do not appear in the lists were not lemmatised, and were treated as separate words. Thus, UNHAPPINESS contains two level 3 affixes, UN- and –NESS, and is lemmatised as HAPPY. HAPPY is a level 1 words, and therefore UNHAPPINESS scores zero points. In contrast, LAUGHABLE contains an affix –ABLE which is not included in the level 2 or 3 lists. LAUGHABLE is therefore not lemmatised as LAUGH. Although LAUGH is a level 1 word, LAUGHABLE is not, and it therefore scores one point for the subject.

Level 2

Inflectional suffixes:

- Plural
- 3rd person singular present tense
- past tense
- past participle
- -ing
- comparative
- superlative
- possessive

Level 3

Most frequent and regular derivational affixes:

- -able not when added to nouns
- -er
- -ish
- -less
- -ly
- -ness
- -th cardinal – ordinal only
- -y adjectives from nouns
- non-
- un-

Appendix 3 Spoken and Written responses from subject 1 (Chapter 4)

	Written responses	Spoken responses
1. attack	*missile game war*	*iraq side injure*
2. board	*chance free summer Sunday*	*black wood international*
3. close	*window valve car door*	*door text shop book*
4. cloth	*shirt skirt pants*	*button closet*
5. dig	*over drill dog*	*drop*
6. dirty	*boring oil garbage toilet*	*hurry child pool cloth*
7. disease	*soil sahara mongol gobi sand*	*doctor medicine*
8. experience	*spectacular worst emotional*	*first*
9. fruit	*apple banana melon grape*	*sweet delicious fever bad*
10. furniture	*drawer sofa*	*cloth house chair bed*
11. habit	*forget tobacco beer*	*nail*
12. hold	*launch up parts*	*have bag wrestling*
13. hope	*on space station shuttle*	*holiday happy*
14. kick	*power mind soccer*	*ball desk*
15. map	*google earth bird train*	*town knowledge*
16. obey	*law rule*	*army*
17. pot	*plant garden hot*	*plant flower*
18. potato	*black chips boy*	*german flew*
19. real	*germany inca documentary news*	*world money life myself*
20. rest	*part summer winter*	*bench chair shoulder sleep*
21. rice	*japan china india curry*	*japanese meal fried*
22. science	*technology electron atom*	*chemical experiment*
23. seat	*train car economy class*	*old train*
24. spell	*English difficult test russian*	*write remember miss*
25. substance	*Atom chemical oxygen nitrogen*	*bench*
26. stupid	*bad pig*	*pig me*
27. television	*electronics broadcasting news*	*funny interesting stupid*
28. tooth	*mouse bit doctor*	*dirty meal*
29. trade	*country container company*	*international import export*
30. window	*open wind close microsoft*	*goods windows open*

Appendix 4 Cue words and sample responses taken at test time one and two subject 1 (Chapter 5)

Lexorig subject 1	TEST TIME ONE	TEST TIME TWO
1. attack	*tiger, terrorism, cat, my leg*	*kick, war, terrorism, 911*
2. board	*ship, plane, trip, sea*	*snow, surf, magnet, chess*
3. close	*friend, cleaning shop, my home,*	*ten, gate, factory, supermarket*
4. cloth	*warm, coat, t-shirt, pants*	*fur, duster, sock, white*
5. dig	*deep, mad, tired, grass*	*dog, sandbox, kindergartner,*
6. dirty	*toilet, small children, my home,*	*scoop, garbage, black, smell,*
7. disease	*bug, cold, dislike, sleep, drag*	*cancer, infectious, hiv, death*
8. experience	*good, education, money, tennis*	*practice, subjective, volunteer*
9. fruit	*orange, apple, grape, melon*	*juicy, mandarin, apple, banana*
10. furniture	*dining board, desk, table, bed*	*wood, steel, table, chair*
11. habit	*tennis, ski, cat, fun*	*sparrow, ski, skate, tennis*
12. hold	*heavy, important, bag, cat*	*dumb-bell, muscle, detergent,*
13. hope	*dream, expect, bright, effort*	*chest, star, 18, despair*
14. kick	*soccer, ball, game, fight*	*can, soccer, football, goal*
15. map	*osaka, trip, street, home*	*geography, distance, navigate*
16. obey	*pain, order, escape, jail*	*egypt, order, firm*
17. pot	*hot, coffee, tea, cocoa*	*water, boil, fire,*
18. potato	*curry, german, peeler, mashed*	*ham, onion, salad, vegetable*
19. real	*friend, believe, always one,*	*story, picture, game, society*
20. rest	*caprice relief, bed, sofa, home*	*sofa, work, sleep*
21. rice	*white, laver, hot, soft*	*boil, white, wash, scoop*
22. science	*study, maniac, experiment*	*evidence, magazine, , report*
23. seat	*sit, relax, sofa, study*	*music, cold, cushion, toilet*
24. spell	*wrong, alphabet, english, kanji*	*write, pen, eraser, memorize*
25. substance	*desk, chair, book, black board*	*chemical, physical, material,*
26. stupid	*me, hate, sad, irritate*	*real young, pooh, baby, stop*
27. television	*program, pc, game, radio*	*antenna, digital, analogue, live,*
28. tooth	*itch, white, smell, paste*	*white, diamond, paste, brush*
29. trade	*severe, red, buy, foreign,*	*sports, owner, money, move*
30. window	*countries open, cold, curtain,*	*bird, rain, outside, curtain*

JC1k subject 1	TEST TIME ONE	TEST TIME TWO
1. away	*go, cold, drink, bicycle*	*cold, snow, rain, cat*
2. blow	*cold, wind, walk, away*	*wind, wig, bird, hat*
3. brush	*teeth, socks, paste, cleaning*	*wash, cup, car, clean*
4. chance	*try, shot, tennis, decide*	*good, bad, music, get*
5. common	*friend, knowledge, sense, fun*	*sense, general, floyd, ethics*
6. dance	*ballet, beautiful, swan, lake*	*music, revolution, arrow, game*
7. district	*lake, large, quiet street*	*straight, right, left, corner*
8. ever	*continue, before, love, unchanged,*	*now, before, after, future*
9. famous	*talent,*	*star, talent, rich, tv*
10. flag	*wave, red, white, japan*	*blow, star, line, colorful*
11. get	*money, lover, bonus, good*	*money, present, give, birthday*
12. head	*hair, face, cap, large*	*bold, hair, important, hard*
13. insect	*spider, dirty, ugly, bug*	*hercules, horn, brown, disease*
14. knee	*hit, bruise, sit, pad*	*hit, supporter, pain, bruise*
15. list	*watch, tennis, slice, cut*	*cut, watch, snap, move*
16. mat	*bath, front door,*	*bath, wrestling, exercise, jump*
17. mountain	*climb, mountain, ski mother, king, rule,*	*white, green, high, oxygen*
18. oil	*check, uniform, life, dull*	*burn, fire, gas, expensive*
19. pattern	*bike, scary,*	*one, repeat, check, print*
20. policeman	*white, panda*	*uniform, pocketbook, cherry,*
21. public	*library, free, useful, convenient*	*toilet, health, money, park*
22. religion	*christian, college, buddhism,slam*	*believe, white, flower, memorial*
23. secret	*game, cold, hard, back*	*key, play, pet, scolded*
24. shirt	*skirt, summer, vacation, hair*	*pants, skirt, cut, hair*
25. sorry	*bad, excuse, angry, plea*	*welcome, apologize, angry,quarrel*
26. smell	*stink, rotten, socks, disgusting*	*cookie, dust, harm, delicious*
27. spirit	*quickly, german, pen, pencil*	*samurai, japanese, soccer, body*
28. surprise	*shame, study, effort, god*	*party, birthday, home, cake*
29. telephone	*mother, handy, call, friend*	*internet, adsl, isdn, cable*
30. tool	*april, let, stupid, dull*	*wood, fight, knife, mountain*

JC2k Subject 1	TEST TIME ONE	TEST TIME TWO
1. affect	*movie, cat, move, love*	*pheromone, ladies, disease, bed*
2. area	*territory, cat, radio wave, move*	*territory, town, osaka, kansai*
3. balance	*ball, seesaw, instable, umbrella*	*foot, ball, board, arm*
4. boundary	*korea, sea, air*	*sea, soccer, goal, out*
5. cement	*hard, white, slick, skate*	*road, stick, foot, flat*
6. comment	*severe, tennis, advice, tv*	*retire, live door, tv, test*
7. connect	*consent, tv, wreaker, defective*	*cord, cut, return, send*
8. court	*tennis, judge, clay, run*	*lawyer, tv, grass, bound*
9. degree	*temperature, study, chemistry, hard*	*temperature, heavy, right*
10. dismiss	*restriction, no pay, homeless, park*	*firm, occupation, homeless, manager*
11. energy	*power, plant, nature, work*	*drink, eat, plant, sun*
12. extreme	*tired, science, busy, changing*	*dangerous, degree, little, short*
13. flow	*water, bath, pool, electricity*	*water, snow, slope, river*
14. goal	*finish, happy, malaise, attain*	*run, soccer, line, pistol*
15. hook	*cherished, hand, arm, lose*	*hanger, clip, shoot, key*
16. index	*library, card, finger, note*	*hand, finger, marry, point*
17. just	*late, long, right, on line*	*safe, out, wait, minute*
18. load	*heavy, freight, car, track*	*heavy, track, gram, ship*
19. memory	*lose, study, bad, full*	*brain, heart, money, card*
20. oblige	*present, job, school, uniform*	*volunteer, order, teacher, student*
21. pain	*bruise, cut, mental, sad*	*hit, chest, death, cut*
22. point	*finger, black board, win, game*	*arrow, hand, stock, wood*
23. profession	*doctor, teacher, nurse, scientist*	*money, nurse, doctor, teacher*
24. reaction	*fast, run, friend, cool*	*over, angry, surprised, tv*
25. research	*chemistry, hard, internet, pc*	*science, evidence, report, experiment*
26. sale	*run, midori, super market, monday*	*cd, time, monthly, dash*
27. ship	*sea, wave, titanic, awaji island*	*sea, wave, yamato, fish*
28. sport	*soccer, volley, news, skate*	*run, muscle, sun, ski*
29. suit	*job, interview, ceremony, black*	*black, pants, shirt, necktie*
30. tight	*rope, pants, schedule, choke*	*cute, pants, fat, neck*

Appendix 5 The Productive Levels Test (Version C)

LEVELS TEST OF PRODUCTIVE VOCABULARY: Parallel Version 1 (Version C)

Complete the underlined words. The example has been done for you.

He was riding a bic*ycle*.

The 2000-word level

1. I'm glad we had this opp__ to talk.
2. There are a doz___ eggs in the basket.
3. Every working person must pay income t___ .
4. The pirates buried the trea___ on a desert island.
5. Her beauty and cha___ had a powerful effect on men.
6. La__ of rain led to a shortage of water in the city.
7. He takes cr___ and sugar in his coffee.
8. The rich man died and left all his we___ to his son.
9. Pup___ must hand in their papers by the end of the week.
10. This sweater is too tight. It needs to be stret___ .
11. Ann intro___ her boyfriend to her mother.
12. Teenagers often adm ___ and worship pop singers.
13. If you blow up that balloon any more it will bur___ .
14. In order to be accepted into the university, he had to impr___ his grades.
15. The telegram was deli___ two hours after it had been sent.
16. The differences were so sl___ that they went unnoticed.
17. The dress you're wearing is lov___ .
18. He wasn't very popu___ when he was a teenager, but he has many friends now.

The 3000-word level

1. He has a successful car____ as a lawyer.
2. The thieves threw ac____ in his face and made him blind
3. To improve the country's economy, the government decided on economic ref___ .
4. She wore a beautiful green go____ to the ball.
5. The government tried to protect the country's industry by reducing the Imp____ of cheap goods.
6. The children's games were funny at first, but finally got on the parents' ner____ .
7. The lawyer gave some wise coun____ to his client.
8. Many people in England mow the la____ of their houses on Sunday morning.
9. The farmer sells the eggs that his he___ lays.
10. Sudden noises at night sca____ me a lot.
11. France was proc____ a republic in the 18th century.
12. Many people are inj____ in road accidents every year.
13. Suddenly he was thru____ into the dark room.
14. He perc___ a light at the end of the tunnel.
15. Children are not independent. They are att ____to their parents.
16. She showed off her sle____ figure in a long narrow dress.
17. She has been changing partners often because she cannot have a sta____ relationship with one person.
18. You must wear a bathing suit on a public beach. You're not allowed to be na____ .

The 5000-word level

1. Soldiers usually swear an oa__ of loyalty to their country.
2. The voter placed the ball___ in the box.
3. They keep their valuables in a vau___ at the bank.
4. A bird perched at the window led__ .
5. The kitten is playing with a ball of ya .
6. The thieves have forced an ent___ into the building.
7. The small hill was really a burial mou___ .
8. We decided to celebrate New Year's E___ together.
9. The soldier was asked to choose between infantry and cav__ .
10. This is a complex problem which is difficult to compr___ .
11. The angry crowd sho___ the prisoner as he was leaving the court.
12. Don't pay attention to this rude remark. Just ign___ it.
13. The management held a secret meeting. The issues discussed were not disc___ to the workers.
14. We could hear the sergeant bel___ commands to the troops.
15. The boss got angry with the secretary and it took a lot of tact to soo___ him.
16. We do not have adeq___ information to make a decision.
17. She is not a child, but a mat____ woman. She can make her own decisions.
18. The prisoner was put in soli___ confinement.

The University Word List level

1. There has been a recent tr____ among prosperous families towards a smaller number of children.
2. The ar____ of his office is 25 square meters.
3. Phil____ examines the meaning of life.
4. According to the communist doc__ , workers should rule the world.
5. Spending many years together deepened their inti____ .
6. He usually read the sport sec___ of the newspaper first.
7. Because of the doctors' strike the cli___ is closed today.
8. There are several misprints on each page of this te___ .
9. The suspect had both opportunity and mot___ to commit the murder.
10. They insp___ all products before sending them out to stores.
11. A considerable amount of evidence was accum__ during the investigation.
12. The victim's shirt was satu___ with blood.
13. He is irresponsible. You cannot re___ on him for help.
14. It's impossible to eva___ these results without knowing about the research methods that were used.
15. He finally att___a position of power in the company.
16. The story tells us about a crime and subs___ punishment.
17. In a hom___ class all students are of a similar proficiency.
18. The urge to survive is inh___ in all creatures.

The 10 000-word level

1. The baby is wet. Her dia___ needs changing.
2. The prisoner was released on par____ .
3. Second year University students in the US are called soph___ .
4. Her favorite flowers were or____ .
5. The insect causes damage to plants by its toxic sec___ .
6. The evac___ of the building saved many lives.
7. For many people, wealth is a prospect of unimaginable felic___ .
8. She found herself in a pred____ without any hope for a solution.
9. The deac___ helped with the care of the poor of the parish.
10. The hurricane whi___ along the coast.
11. Some coal was still smol____ among the ashes.
12. The dead bodies were muti____ beyond recognition.
13. She was sitting on a balcony and bas____ in the sun.
14. For years waves of invaders pill___ towns along the coast.
15. The rescue attempt could not proceed quickly. It was imp____ by bad weather.
16. I wouldn't hire him. He is unmotivated and indo___ .
17. Computers have made typewriters old-fashioned and obs___ .
18. Watch out for his wil____ tricks.

Appendix 6 The JACET8000 first thousand words

a	arm	body	clearly	death
ability	army	book	climb	decide
able	around	both	close	decision
about	arrive	box	club	deep
above	art	boy	cold	degree
accept	artist	brain	college	demand
accident	as	branch	color	depend
achieve	ask	break	come	describe
across	at	bright	common	design
act	attack	bring	communication	develop
action	attempt	brother	community	development
activity	attention	build	company	die
actually	attitude	building	compare	difference
add	audience	bus	complete	different
address	average	business	completely	difficult
adult	avoid	but	computer	difficulty
affect	aware	buy	concern	dinner
afraid	away	by	concerned	direction
after	baby	call	condition	director
afternoon	back	can	consider	discover
again	bad	capital	contact	discuss
against	bag	captain	contain	discussion
age	ball	car	continue	disease
ago	bank	card	control	distance
agree	base	care	conversation	do
ahead	basic	career	corner	doctor
aid	be	carefully	cost	dog
air	bear	carry	could	door
all	beat	case	count	doubt
allow	beautiful	cat	country	down
almost	because	catch	couple	draw
alone	become	cause	course	dream
along	bed	cell	court	dress
already	before	center	cover	drink
also	begin	central	create	drive
although	beginning	century	cross	driver
always	behavior	certain	crowd	drop
among	behind	certainly	cry	dry
amount	being	chair	culture	during
an	believe	challenge	cup	each
and	below	chance	cut	early
animal	best	change	dad	earth
another	better	character	damage	easily
answer	between	check	dance	easy
any	beyond	child	danger	eat
anyone	big	choice	dark	economic
anything	bird	choose	date	education
anyway	black	church	daughter	effect
appear	blood	city	day	effort
approach	blue	class	dead	either
area	board	clean	deal	else
argue	boat	clear	dear	encourage
				end
				energy

enjoy	finish	hand	include	leg
enough	fire	happen	increase	less
enter	fish	happy	indeed	let
environment	floor	hard	individual	letter
escape	flower	hardly	industry	level
especially	fly	have	influence	lie
even	follow	he	information	life
evening	following	head	inside	lift
event	food	health	instead	light
ever	foot	hear	interest	like
every	for	heart	interested	likely
everybody	force	heat	interesting	limit
everyone	foreign	heavy	international	line
everything	forest	help	into	list
exactly	forget	her	introduce	listen
example	form	here	involve	little
except	former	herself	island	live
exchange	forward	high	issue	local
exercise	free	hill	it	long
exist	freedom	him	its	look
expect	friend	himself	itself	long
experience	from	his	job	lose
explain	front	history	join	loss
express	full	hit	judge	lot
expression	future	hold	jump	love
eye	game	hole	just	low
face	garden	home	keep	machine
fact	gas	hope	kid	main
fail	general	horse	kill	major
fall	get	hospital	kind	make
family	girl	hot	king	male
famous	give	hotel	kitchen	man
far	glass	hour	know	manage
farm	go	house	knowledge	manager
fast	goal	how	lack	many
father	gold	however	lady	mark
fear	good	huge	land	market
feel	government	human	language	marry
feeling	great	hurt	large	material
few	green	husband	last	matter
field	ground	I	late	may
fight	group	ice	laugh	maybe
figure	grow	idea	law	me
fill	growth	if	lay	meal
film	guess	image	lead	mean
final	guide	imagine	leader	meaning
finally	gun	immediately	learn	measure
find	guy	important	least	meet
fine	hair	improve	leave	meeting
finger	half	in	left	member

memory	nor	pattern	prove	rise
mention	normal	pay	provide	river
message	north	peace	public	road
method	not	people	publish	rock
middle	note	percent	pull	role
might	nothing	performance	purpose	room
mile	notice	perhaps	push	round
mind	now	period	put	rule
mine	number	person	quality	run
minute	object	personal	question	safe
miss	occur	phone	quickly	sale
mistake	of	pick	quiet	same
model	off	picture	quite	save
modern	offer	piece	race	say
mom	office	place	radio	scene
moment	officer	plan	rain	school
money	official	plant	raise	science
month	often	play	rate	scientist
more	oh	player	rather	sea
morning	oil	please	reach	search
most	old	point	read	season
mother	on	police	ready	seat
mountain	once	political	real	secret
mouth	only	poor	realize	see
move	open	popular	really	seem
movement	operation	population	reason	sell
movie	opinion	position	receive	send
much	opportunity	possible	recent	sense
music	or	pound	recently	sentence
must	order	power	record	separate
my	other	practice	red	series
myself	our	prepare	reduce	serious
name	out	present	refuse	serve
nation	outside	president	relationship	service
national	over	press	remain	set
natural	own	pressure	remember	several
nature	page	pretty	reply	shake
near	pain	price	report	shall
nearly	paint	private	represent	shape
necessary	painting	probably	require	share
need	paper	problem	research	she
never	parent	process	resource	ship
new	park	produce	respect	shop
news	part	product	rest	short
newspaper	particular	professional	result	should
next	particularly	program	return	shoulder
nice	party	progress	rich	shout
night	pass	project	ride	show
no	past	promise	right	side
nobody	patient	protect	ring	sight

sign				
similar	stage	telephone	under	whole
simple	stand	television	understand	whose
simply	standard	tell	understanding	why
since	star	tend	united	wide
sing	stare	test	university	wife
single	start	than	until	wild
sir	state	thank	up	will
sister	station	that	upon	win
sit	stay	the	us	wind
situation	step	their	use	window
size	stick	them	usually	winter
skill	still	themselves	value	wish
skin	stone	then	variety	with
sky	stop	theory	various	within
sleep	store	there	very	without
slow	story	therefore	video	woman
slowly	straight	these	view	wonder
small	strange	they	village	wonderful
smile	street	thing	visit	wood
so	stress	think	voice	word
social	strike	this	vote	work
society	strong	those	wait	worker
some	struggle	though	walk	world
someone	student	thought	wall	worry
something	study	through	want	worth
sometimes	style	throughout	war	would
son	subject	throw	warm	write
song	success	thus	wash	writer
soon	successful	time	waste	wrong
sorry	such	to	watch	year
sort	suddenly	today	water	yes
sound	suffer	together	wave	yesterday
source	suggest	tomorrow	way	yet
south	summer	too	we	you
space	sun	top	wear	young
speak	supply	total	weak	your
speaker	support	touch	welcome	yourself
special	suppose	toward	well	
species	sure	town	west	
speech	surface	trade	western	
speed	surprise	traditional	what	
spend	survive	train	whatever	
spirit	system	training	when	
sport	table	travel	where	
spread	take	treat	whether	
spring	talk	tree	which	
	tea	trip	while	
	teach	trouble	white	
	teacher	true	who	
	team	trust		
	tear	truth		
	technology	try		
		turn		
		type		

Appendix 7 The General Service List first thousand words

able	as	business	council	dry
about	ask	but	count	due
above	associate	buy	country	duty
accept	at	by	course	each
accord	attack	call	court	ear
account	attempt	can	cover	early
accountable	average	capital	cross	earth
across	away	captain	crowd	east
act	back	car	crown	easy
active	bad	care	cry	eat
actor	ball	carry	current	effect
actress	bank	case	cut	efficient
actual	bar	castle	danger	effort
add	base	catch	dark	egg
address	battle	cause	date	eight
admit	be	centre	daughter	either
adopt	bear	certain	day	elect
advance	beauty	chance	dead	eleven
advantage	because	change	deal	else
adventure	become	character	dear	empire
affair	bed	charge	decide	employ
after	before	chief	declare	end
again	begin	child	deep	enemy
against	behind	choose	defeat	english
age	believe	church	degree	enjoy
agent	belong	circle	demand	enough
ago	below	city	department	enter
agree	beneath	claim	depend	equal
air	beside	class	describe	escape
all	best	clear	desert	even
allow	between	close	desire	evening
almost	beyond	cloud	destroy	event
alone	big	coal	detail	ever
along	bill	coast	determine	every
already	bird	coin	develop	example
also	black	cold	die	except
although	blood	college	difference	exchange
always	blow	colony	difficult	exercise
among	blue	colour	direct	exist
amount	board	come	discover	expect
ancient	boat	command	distance	expense
and	body	committee	distinguish	experience
animal	book	common	district	experiment
another	both	company	divide	explain
answer	box	complete	do	express
any	boy	concern	doctor	extend
appear	branch	condition	dog	eye
apply	bread	consider	dollar	face
appoint	break	contain	door	fact
arise	bridge	content	doubt	factory
arm	bright	continue	down	fail
army	bring	control	draw	fair
around	broad	corn	dream	faith
arrive	brother	cost	dress	fall
art	build	cotton	drink	familiar
article	burn	could	drive	family

famous	get	indeed	literature	most
far	gift	independent	little	mother
farm	girl	industry	live	motor
fast	give	influence	local	mountain
father	glad	instead	long	mouth
favour	glass	interest	look	move
fear	go	into	lord	mrs
feel	god	introduce	lose	much
fellow	gold	iron	loss	music
few	good	it	love	must
field	great	join	low	name
fight	green	joint	machine	nation
figure	ground	jointed	main	native
fill	group	joy	make	nature
find	grow	judge	man	near
fine	half	just	manner	necessary
finish	hand	justice	manufacture	necessity
fire	hang	keep	many	need
first	happen	kill	mark	neighbour
fish	happy	kind	market	neither
fit	hard	king	marry	never
five	hardly	know	mass	new
fix	have	lack	master	news
floor	he	lady	material	newspaper
flow	head	lake	matter	next
flower	hear	land	maybe	night
fly	heart	language	mean	nine
follow	heat	large	measure	no
food	heaven	last	meet	noble
for	heavy	late	member	none
force	help	latter	memory	nor
foreign	here	laugh	mention	north
forest	high	laughter	mere	not
forget	hill	law	*met al*	note
form	history	lay	middle	notice
former	hold	lead	might	now
forth	home	learn	mile	number
fortune	honour	leave	milk	numerical
four	hope	left	million	numerous
free	horse	length	mind	object
fresh	hot	less	miner	observe
friday	hour	let	minister	occasion
friend	house	letter	minute	of
from	how	level	miss	off
front	however	library	mister	offer
full	human	lie	modern	office
furnish	hundred	life	moment	official
future	husband	lift	monday	often
gain	idea	light	money	oh
game	if	like	month	oil
garden	ill	likely	moon	old
gas	important	limit	moral	on
gate	in	line	more	once
gather	inch	lip	moreover	one
general	include	listen	morning	only
gentle	increase			

open	pretty	reply	several	spring
operate	prevent	report	shadow	square
opinion	price	represent	shake	stage
opportunity	private	republic	shall	stand
or	problem	reserve	shape	standard
order	produce	respect	share	star
ordinary	product	rest	she	start
organize	profit	result	shine	state
other	progress	return	ship	station
otherwise	promise	rich	shoot	stay
ought	proof	ride	shore	steel
out	proper	right	short	step
over	property	ring	should	still
owe	propose	rise	shoulder	stock
own	protect	river	show	stone
page	prove	road	side	stop
paint	provide	rock	sight	store
paper	provision	roll	sign	story
part	public	room	silence	strange
particular	pull	rough	silver	stream
party	purpose	round	simple	street
pass	put	royal	since	strength
past	quality	rule	sing	strike
pay	quantity	run	single	strong
peace	quarter	safe	sir	struggle
people	queen	sail	sister	student
per	question	sale	sit	study
perhaps	quite	salt	situation	subject
permit	race	same	six	substance
person	raise	saturday	size	succeed
picture	rank	save	sky	such
piece	rate	say	sleep	suffer
place	rather	scarce	small	suggest
plain	reach	scene	smile	summer
plan	read	school	snow	sun
plant	ready	science	so	sunday
play	real	sea	social	supply
please	realise	season	society	support
point	really	seat	soft	suppose
political	reason	second	soldier	sure
poor	receipt	secret	some	surface
popular	receive	secretary	son	surprise
population	recent	see	soon	surround
position	recognize	seem	sort	sweet
possess	record	sell	soul	sword
possible	red	send	sound	system
post	reduce	sense	south	table
pound	refuse	sensitive	space	take
poverty	regard	separate	speak	talk
power	relation	serious	special	tax
prepare	relative	serve	speed	teach
present	religion	service	spend	tear
president	remain	set	spirit	tell
press	remark	settle	spite	temple
pressure	remember	seven	spot	ten
			spread	

term	use	within		
test	usual	without		
than	valley	woman		
the	value	wonder		
then	variety	wood		
there	various	word		
therefore	very	work		
they	vessel	world		
thing	victory	worth		
think	view	would		
thirteen	village	wound		
thirty	virtue	write		
this	visit	wrong		
though	voice	year		
thousand	vote	yes		
three	wage	yesterday		
through	wait	yet		
throw	walk	yield		
thursday	wall	you		
thus	want	young		
till	war	youth		
time	watch			
to	water			
today	wave			
together	way			
ton	we			
too	wealth			
top	wear			
total	wednesday			
touch	week			
toward	welcome			
town	well			
trade	west			
train	western			
travel	what			
tree	when			
trouble	where			
true	whether			
trust	which			
try	while			
tuesday	white			
turn	who			
twelve	whole			
twenty	why			
two	wide			
type	wife			
under	wild			
understand	will			
union	win			
unite	wind			
university	window			
unless	winter			
until	wise			
up	wish			
upon	with			

Bibliography

Adolphs.S., and Schmitt. N. (2003). Lexical Coverage of Spoken Discourse. *Applied Linguistics,* 24, 425-438.

Aizawa, K. (2006). Rethinking Frequency Markers For English-Japanese Dictionaries. In M Murata., K. Minamiide, Y. Tono and Ishikawa. S. (Eds.) *English Lexicography in Japan*. Tokyo: Taishukan.

Anderson, R. and Freebody. P. (1981). Vocabulary Knowledge. In J.T. Guthrie. (Ed.), *Comprehension and teaching: Research reviews*, 77-117. Newark, DE: International Reading Association.

Anderson, R. C., and Nagy, W. (1992).The Vocabulary Conundrum. *American Educator,* 16, 4, 1418, 44-47.

Baba, K. (2002). Test Review: Lex30. *Language Testing Update,* 32, 68-71.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

Bachman, L. F. (2004). *Statistical Analyses for Language Assessment.* Cambridge: Cambridge University Press.

Bachman, L. F., and Palmer, A. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Bauer, L., and Nation, I.S.P. (1993). Word Families. *International Journal of Lexicography* 6, 253-279.

Beheydt, L. (1987). The Semantization Of Vocabulary In Foreign Language Learning. *System,* 15 (1), 55-67.

Bogaards, P. (2001). Lexical Units And The Learning Of Foreign Vocabulary. *Studies in Second Language Acquisition,* 23, *3,* 321-343.

Blot, K. J., Zárate, M. A., and Paulus, P. B. (2003). Code-Switching across Brainstorming Sessions: Implications for the Revised Hierarchical Model of Bilingual Language Processing. *Experimental Psychology*, 50, 171-183.

Canale, M. and Swain, M. (1980). Theoretical Bases Of Communicative Approaches To Second Language Teaching And Testing. *Applied Linguistics* 1, 1-47.

Chapelle, C.A. (1998). Construct Definition And Validity Inquiry In SLA Research. In Bachman, L.F. and Cohen, A.D. (Eds.), *Interfaces Between Second Language Acquisition And Language Testing Research.* Cambridge: Cambridge University Press, 32–70.

Cobb,T. *Web Vocabprofile* [accessed October 2005 to January 2010] from http://www.lextutor.ca/vp/], an adaptation of Heatley and Nation's (1994) *Range.*

Cohen, A. (1994). *Assessing Language Ability In The Classroom.* Boston: Heinle and Heinle Publishers.

Cronbach, L. J. (1942). An Analysis Of Techniques For Diagnostic Vocabulary Testing. *Journal of Educational Research,* 36, 3, 206-217.

Cruse, D.A. (1986). *Lexical Semantics*. Cambridge: Cambridge University Press.

Cummins, J. (2000). *Language, Power, And Pedagogy: Bilingual Children In The Crossfire.* Clevedon, England: Multilingual Matters.

De Groot, A. M. B. (1992). Bilingual Lexical Representation: A Closer Look At Conceptual Representations. In R. Frost and L. Katz (Eds.), *Orthography, Phonology, Morphology, And Meaning,* 389-412. Amsterdam: North-Holland.

De Groot, A.M.B. (1993). Word-Type Effect Effects In Bilingual Processing Tasks: Support For A Mixed Representational System. In Schreuder, B.W. (Ed.), *The Bilingual Lexicon,* 27–51, John Benjamins.

De Groot, A. M. B., Dannenburg, L., and Van Hell, J.G.(1994). Forward And Backward Word Translation By Bilinguals. *Journal of Memory and Language,* 33, 600-629.

De Groot, A.M.B. and Hoeks, J.C. (1995). The Development Of Bilingual Memory: Evidence From Word Translation By Trilinguals. *LanguageLearning,* 45, 683–724.

Den Dulk, J.J. (1985). *Productive Vocabulary And The Word Association Test.* Unpublished master's thesis. University of Utrecht, Utrecht.

Dolch, E.W. and Leeds, D. (1953). Vocabulary Tests And Depth Of Meaning. *Journal of Educational Research,* 47, 181-189.

Faerch, C., Haastrup. K., and Phillipson, R. (1984). *Learner Language And Language Learning*. Clevedon, England, Multilingual matters.

Fitzpatrick, T. (2000). Using Word Association Techniques to Measure Productive Vocabulary in a Second Language. *Language Testing Update,* 32, 68-71.

Fitzpatrick, T. (2003). *Eliciting And Measuring Productive Vocabulary Using Word Association Techniques And Frequency Bands.* PhD Thesis, University of Wales, Swansea.

Fitzpatrick, T. (2004). Measuring Short-term Changes in the Lexicon of L2 Learners. Poster presented at the *14 EUROSLA Conference*. University of the Basque Country, San Sebastián.

Fitzpatrick. T. (2007). Productive Vocabulary Tests And The Search For Concurrent Validity". In Daller, H., Milton, J. and J. Treffers-Daller. (Eds.) *Modelling and Assessing Vocabulary Knowledge.* Cambridge: Cambridge University Press.

Fitzpatrick. T. and Meara P. (2004). Exploring The Validity Of A Test Of Productive Vocabulary. *VIAL* 1, 55-74.

Gipps, C.V.(1994). *Beyond Testing: Towards a Theory of Educational Assessment.* The Falmer Press: London, Philadelphia.

Heatley, A. and Nation, P. (1994). *Range*. Victoria University of Wellington, NZ. Computer program, available at http://www.vuw.ac.nz/lals/.

Heatley, A., and Nation, I.S.P., (1998*) VocabProfile and Range*. School of Linguistics and Applied Language Studies. Victoria University of Wellington, Wellington, New Zealand.

Henriksen, B. (1996). *Semantisation, Retention and accessibility: Key Concepts in Vocabulary Learning*. Paper presented at the AILA Congress, Jyvaskyla, Finland. August 1996.

Henriksen, B. (1999). Three Dimensions Of Vocabulary Development. *Studies in Second Language Acquisition* , 21, 303—317.

Horst, M., Cobb, T., and Meara, P. (1998). Beyond A Clockwork Orange: Acquiring Second Language Vocabulary Through Reading. *Reading in a Foreign Language,* 11, 207–223.

Ishikawa, S., Uemura T., Kaneda, M., Shmizu, S., Sugimori, N. Tono, Y. Mochizuki, M and M. Murata. (2003). *JACET 8000: JACET List of 8000 Basic Words.* Tokyo: JACET.

Ishikawa, S. and Uemura, T. (2004). JACET 8000 and Asian TEFL Initiative. *The Journal of Asia TEFL*, 1(1), 333-347.

Japan Association of College English Teachers *(2003). JACET List of 8000 Basic Words. Tokyo: JACET.*

Jiménez Catalán, R. M. and Moreno Espinosa. S. (2005).Using Lex30 To Measure The L2 Productive Vocabulary Of Spanish Primary Learners Of EFL. *VIAL*, 2, 27-44.

Kiss, G.R., Armstrong, C.A., and Milroy, R. (1973). An Associative Thesaurus of English. *EP Microfilms,* Wakefield.

Kroll, J.F. and Stewart, E. (1994). Category Interference In Translation And Picture Naming: Evidence For Asymmetric Connections Between Bilingual Memory Representations. *Journal of Memory and Language,* 33, 149–174.

Kroll, J.F., Michael, E., Tokowicz.N., and Dufour, R. (2002). The Development Of Lexical Fluency In A Second Language. *Second Language Research,* 18, 2, 137–171.

Kruse, H., Pankhurst, J. and Sharwood Smith. M. (1987). A Multiple Word Association Probe In Second Language Acquisition Research. *Studies in Second Language Acquisition,* 9, 2, 141-154.

Larsen-Freeman, D. and Cameron, L. (2008). *Complex Systems and Applied Linguistics.* Oxford: Oxford University Press.

Laufer, B. (1989). What Percentage Of Text Lexis Is Essential For Comprehension? In Ch. Lauren and M. Nordman. (Eds.) *Special Language: From Humans Thinking To Thinking Machines*, 316-323. Multilingual Matters.

Laufer, B. (1991). Knowing A Word: What Is So Difficult About It? *English Teachers' Journal, 42* (May), 82-88.

Laufer, B. (1992). How Much Lexis Is Necessary For Reading Comprehension? In P.J. Arnaud. and H. Boint. (Eds.), *Vocabulary And Applied Linguistics*,126-132. London: Macmillan.

Laufer, B. (1994). The Lexical Profile Of Second Language Writing: Does It Change Over Time? *RELC Journal,* 25,2, 21-33.

Laufer, B. (1995). Beyond 2000 - A Measure Of Productive Lexicon In A Second Language. In *The Current State of Interlanguage*, (Eds.) L. Eubank, M. Sharwood-Smith, L.Selinker. Benjamins. 265-272. Studies In Honor Of William E. Rutherford, 265-272. Amsterdam: John Benjamins.

Laufer, B. (1998). The Development Of Passive And Active Vocabulary In A Second Language: Same Or Different? *Applied Linguistics,* 12, 255-271.

Laufer, B. (2001). Quantitative Evaluation Of Vocabulary: How It Can Be Done And What It Is Good For? In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. McNamara and K. O'Loughlin (Eds.) *Experimenting With Uncertainty: Essays In Honour Of Alan Davies*, 241-250. Cambridge: Cambridge University Press.

Laufer, B. (2005). Lexical Frequency Profiles: From Monte Carlo to the Real World. A response to Meara. *Applied Linguistics,* 26, 581-587.

Laufer, B., and Nation, I.S.P. (1995). Vocabulary Size And Use: Lexical Richness In L2 Written Production. *Applied Linguistics,* 16, 307-322.

Laufer, B., and Paribakht. T. S. (1998). The Relationship Between Passive And Active Vocabularies: Effects Of Language Learning Context. *Language Learning,* 48, 3, 365-391.

Laufer, B., and Nation, P. (1999). A Vocabulary-Size Test Of Controlled Productive Ability. *Language Testing,* 16, 1, 33-51.

Laufer, B., Elder, C., Hill, K., and Congdon, P. (2004). Size And Strength: Do We Need Both To Measure Vocabulary Knowledge? *Language Testing,* 21, 202-226.

Madsen, H. S. (1983). *Techniques in testing*. New York: Oxford University Press.

Meara, P. (1983). Word Associations In A Foreign Language. *Nottingham Linguistics Circular,* 11, 29-38.

Meara, P. (1990). A Note On Passive Vocabulary. *Second Language Research,* 6, 2, 150–154.

Meara, P. (1994). The Complexities Of Simple Vocabulary Tests. In F.G. Brinkman, J.A. van der Schee, and M.C. Schouten-van Parreren, (Eds.), *Curriculum Research: Different Disciplines and Common Goals.* Vrije Universiteit, Amsterdam.

Meara, P. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjaer, and J. Williams. (Eds.), *Performance And Competence In Second Language Acquisition*, 35–53. Cambridge: Cambridge University Press.

Meara, P. (2005). Lexical frequency profiles: a Monte Carlo analysis. *Applied linguistics* 26, 32-47.

Meara, P. (2006). Emergent Properties of Multilingual Lexicons. *Applied Linguistics,* 27, 620-644.

Meara, P. and Buxton. B. (1987). An Alternative To Multiple Choice Vocabulary Tests. *Language Testing,* 4, 2, 142-151.

Meara, P., and Jones, G., (1990). Vocabulary Size A Placement Indicator. In P. Grunwell., (Ed.), *Applied Linguistics in Society.* CILT, London pp. 80-87.

Meara, P. and I. Rodriguez Sanchez. (1993). Matrix Models Of Vocabulary Acquisition: An Empirical Assessment. In M. Wesche and T. S. Paribakht (Eds.), *Symposium on Vocabulary Research.* Ottawa: CREAL.

Meara, P. and Fitzpatrick, T. (2000). Lex30: An Improved Method Of Assessing Productive Vocabulary In An L2. *System,* 28, 19-30.

Meara, P. and Milton, J.L. (2002). *X_Lex: The Swansea Vocabulary Levels Test.* Newbury, UK: Express Publishing.

Meara, P., and Fitzpatrick, T. (2004) *Lex Scorer V 0.1. Swansea: Lognostics.*

Meara, P.M. and Wolter, B. (2004) 'V_Links: Beyond Vocabulary Depth', *Angles on the English Speaking World,* 4, 85-96.

Melka, F (1982). Receptive Versus Productive Vocabulary: A Survey. *Interlanguage Studies Bulletin*, 6, 5-33.

Melka, F. (1997). Receptive Vs. Productive Aspects Of Vocabulary. In N. Schmitt and M. McCarthy (Eds.), Vocabulary, Description, Acquisition And Pedagogy, 84–102. New York: Cambridge University Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.), 13-104. New York: Macmillan.

Miller, G A. and Fellbaum. C (1991). Semantic Networks Of English. *Cognition*, 41, 197-229.

Mizumoto, A. and Takeuchi, O. (2009). A Closer Look At The Relationship Between Vocabulary Learning Strategies And The TOEIC Scores. *TOEIC Research Report, 4*, 1-34. Tokyo: IIBC.

Moreno Espinosa, S. (2010). Boys' and Girls' L2 Word Associations. In Jiménez Catalán, R. (Ed.) *Gender Perspectives on Vocabulary in Foreign and Second Languages*. Palgrave Macmillan.

Moreno Espinosa, S. (2009). Young Learners' L2 word association responses in two different learning contexts. In Ruiz de Zarobe, Y. & Jiménez Catalán, R. (Eds.) *Content and Language Integrated Learning in Europe*. Multilingual Matters.

Nation, I. S. P. (1983). Testing And Teaching Vocabulary. *Guidelines, 5,* 12-25.

Nation, I. S. P. (1983). Teaching And Learning Vocabulary. *English Language Institute,* Wellington: University of Wellington.

Nation, I. S. P. (1984). *Vocabulary Lists*. Wellington: Victory University of Wellington, English Language Institute.

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. Boston: Heinle and Heinle.

Nation, I. S. P. (1990b). Measuring Readiness for Simplified Material: A Test of the first 1,000 Words of English. In M.L. Tickoo. (Ed.) *Simplification: Theory and Application. RELC Anthology Series 31.*

Nation, I.S. P., and Wang, K. M. (1999). Graded Readers And Vocabulary. *Reading in a Foreign Language 12*, 2, 355-380.

Nation, I.S.P. (2001). *Learning Vocabulary In Another Language.* Cambridge: Cambridge University Press.

Nation, I.S.P. (2005). *Teaching Vocabulary: Is it a Waste of Learning Time?* Plenary talk given at TESOL Conference, San Antonio, TX.

Palmberg, R. (1987). Patterns Of Vocabulary Development In Foreign-Language Learners. *Studies in Second Language Acquisition,* 9, 201-220.

Palmberg, R. (1989). What Makes A Word English? Swedish Speaking Learners' Feeling Of "Englishness." *AILA Review*, 6, 47–55.

Paribakht, T. S., and Wesche, M. (1993). Reading Comprehension And Second Language Development In A Comprehension-Based ESL Program. *TESL Canada Journal,* 11, 9-29.

Pearson, P. D., Hiebert, E. H., and Kamil, M. L. (2007). Vocabulary Assessment: What We Know and What We Need to Learn. *Reading Research Quarterly*, 42, 2, 282-296.

Potter, M.C., Kwok-Fai, S., Von Eckardt, B., Feldman, L. B. (1984). Lexical And Conceptual Representation In Beginning And Proficient Bilinguals. *Journal of Verbal Learning and Verbal Behavior*, 23, 23–38.

Randall, M. (1980). 'Word Association Behaviour In Learners Of English As A Second Language', *Polyglot* 2, 2, B4-Dl.

Read, J. (1988). Measuring The Vocabulary Knowledge Of Second Language Learners. *RELC Journal* 19, 12–25.

Read, J. (2000). *Assessing Vocabulary Knowledge And Use.* Cambridge: Cambridge University Press.

Read, J. (2004). 'Plumbing the Depths: How should the Construct of Vocabulary Knowledge be Defined?' In P. Bogaards and B. Laufer (Eds.), *Vocabulary In A Second Language*, 209-227. Amsterdam: John Benjamins.

Read, J. and Chapelle, C. A. (2001). A Framework For Second Language Vocabulary Assessment. *Language Testing,* 18, 1, 1-32.

Richards, J. C. (1976).The Role of Vocabulary Teaching. *TESOL Quarterly,* 10, 1, 77-89.

Ringbom, H. (1987*). The Role Of The First Language In Foreign Language Learning.* Clevedon and Philadelphia: Multilingual Matters.

Richards, B.J. and Malvern, D. D. (2007). Validity And Threats To The Validity Of Vocabulary Measurement. In H. Daller, J. Milton, and J. Treffers-Daller. (Eds.) *Modelling and Assessing Vocabulary Knowledge.* Cambridge: Cambridge University Press.

Schmitt, N. (1998). Tracking the Incremental Acquisition of Second Language Vocabulary: A Longitudinal Study. *Language Learning,* 48, 281-317.

Schmitt, N. (1999). The Relationship Between TOEFL Vocabulary Items And Meaning, Association, Collocation And Word-Class Knowledge. *Language Testing,* 16, 189-216.

Schmitt, N. and McCarthy. M. (Eds.). (1997). *Vocabulary*: *Description, Acquisition, and Pedagogy*. Cambridge: Cambridge University Press.

Schmitt, N. and Meara, P. (1997). Researching Vocabulary Through A Word Knowledge Framework. *SSLA,* 20, 17-36.

Schmitt, N., Schmitt, D. and Clapham, C. (2001). Developing And Exploring The Behaviour Of Two New Versions Of The Vocabulary Levels Test. *Language Testing,* 18, 1, 55-88.

Schonell, F. Meddleton, I., Shaw, B., Routh, M., Popham, D., Gill, G., Mackrell, G., and Stephens, C. (1956). *A Study Of The Oral Vocabulary Of Adults.* Brisbane and London: University of Queensland Press / University of London Press.

Spreen, O., and Benton, A. L. (1977). Neurosensory Center Comprehensive Examination for Aphasia. Victoria, B. C.: University of Victoria Neuropsychology Laboratory.

Thorndike, E.L. and Lorge, I. (1944). The Teacher's Word Book Of 30,000 Words. Teachers College, Columbia University, New York.

Van Hell, J. G., and De Groot, A. M. B. (1998). Disentangling Context Availability And Concreteness In Lexical Decision And Word Translation. *The Quarterly Journal of Experimental Psychology,* 51A, 41-63.

Vermeer, A. (2001). Breadth And Depth Of Vocabulary In Relation To L1/L2 Acquisition And Frequency Of Input. *Applied Psycholinguistics,* 22, 217–34.

Waring. R. (1997). A Study Of Receptive And Productive Learning From Word Cards. *Studies in Foreign Languages and Literature*. Notre Dame Seishin University, Okayama, 21, 1, 94-114.

Waring, R., (1999). *The Measurement Of Receptive And Productive Vocabulary*. PhD thesis, University of Wales, Swansea.

Waring, R. and Nation, I.S.P. (2004) Second language reading and incidental vocabulary learning. *Angles on the English Speaking World,* 4, 97-110.

Waring, R., and Takaki, M. (2003). At What Rate Do Learners Learn And Retain New Vocabulary From Reading A Graded Reader? *Reading in a Foreign Language,* 15, 130-163.

Webb, S. (2005). The Effect of Reading and Writing on Word Knowledge. *Studies in Second Language Acquisition,* 27, 33-52.

Webb, S. (2007). The Effects Of Repetition On Vocabulary Knowledge. *Applied Linguistics* 28, 46-65.

Wesche, M. and Paribakht. T. (1996). Assessing Vocabulary Knowledge: Depth Vs. Breadth. *Canadian Modern Language Review*, 53, 1, 13-40.

West, M. (1953). *A General Service List of English Words.* London: Longman, Green and Company.

Wolter, B. (2001). Comparing The L1 And L2 Mental Lexicon: A Depth Of Individual Word Knowledge Model. *Studies in Second Language Acquisition,* 23, 41-69.

Zipf, G. (1935). *The Psychobiology of Language*. Boston: Houghton-Mifflin.